

Development of an Afrikaans test for sentence recognition in noise

by

Marianne Theunissen

APRIL 2008

Supervisor: Dr D Swanepoel
Co-supervisor: Prof J J Hanekom

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE:
M COMMUNICATION PATHOLOGY
IN THE DEPARTMENT OF COMMUNICATION PATHOLOGY
FACULTY OF HUMANITIES
UNIVERSITY OF PRETORIA

TABLE OF CONTENTS

TABLE OF CONTENTS	1
LIST OF FIGURES	4
LIST OF TABLES.....	6
LIST OF APPENDICES	8
LIST OF ABBREVIATIONS	9
1. INTRODUCTION AND ORIENTATION	10
1.1 Introduction.....	10
1.2 Background	11
1.3 Rationale	17
1.4 Problem statement	19
1.5 Definition of terms.....	21
1.6 Outline of chapter contents.....	24
1.7 Conclusion.....	25
1.8 Summary	25
2. METHODOLOGICAL CONSIDERATIONS IN THE DEVELOPMENT OF A SPEECH-IN-NOISE TEST	26
2.1 Introduction.....	26
2.2 Framework for the development of speech-in-noise tests	27
2.3 Stimulus variables for speech-in-noise tests.....	29
2.3.1 Speech material	29
2.3.2 Type of noise.....	37
2.3.3 Speaker variables.....	40
2.4 Presentation variables for speech-in-noise tests	43
2.4.1 Presentation method	43
2.4.2 Auditory transmission channel.....	45
2.5 Subject variables for speech-in-noise tests	48
2.5.1 Peripheral hearing	48
2.5.2 Age	50
2.5.3 Central Nervous System.....	52
2.6 Response variables for speech-in-noise tests	55
2.6.1 Response channel.....	55
2.6.2 Scoring method	56
2.7 Test performance variables for speech-in-noise tests	58
2.7.1 Validity.....	59
2.7.2 Reliability	63
2.7.3 Sensitivity and specificity	71
2.8 Conclusion.....	74
2.9 Summary	78
3. METHODOLOGY	79
3.1 Introduction.....	79
3.2 Aims.....	79
3.3 Research Design	80
3.4 Ethical considerations.....	82
3.4.1 Respect for the privacy of research participants.....	82
3.4.2 Informed consent.....	82
3.4.3 Beneficence and non-maleficance	83
3.4.4 Distributive justice.....	83

3.4.5	Cultural and linguistic diversity	83
3.5	Research Sample	83
3.5.1	Population	84
3.5.2	Selection criteria	84
3.5.3	Selection procedures	86
3.5.4	Description of research sample	89
3.6	Material and apparatus	92
3.6.1	Material and apparatus for subject selection	92
3.6.2	Material and apparatus for data collection, recording and analysis	93
3.7	Procedures: Phase I	96
3.7.1	Compilation of sentences	97
3.7.2	Rating of naturalness	100
3.7.3	Rating of grammatical complexity	101
3.7.4	Recording and editing of material	103
3.8	Procedures: Phase II	104
3.8.1	First equalisation procedure: Selection of equivalent subset of sentences	106
3.8.2	Second equalisation procedure: Selecting sentences with similar intelligibility slopes	109
3.9	Procedures: Phase III	113
3.9.1	Compilation of lists	115
3.9.2	Experimental application of lists	122
3.10	Reliability and Validity	126
3.11	Conclusion	130
3.12	Summary	131
4.	RESULTS	134
4.1	Introduction	134
4.2	Phase I	134
4.2.1	Compilation of sentence material	134
4.2.2	Rating of naturalness	136
4.2.3	Rating of grammatical level	138
4.3	Phase II	139
4.3.1	First equalisation procedure: Selection of equivalent subset of sentences	140
4.3.2	Second equalisation procedure: Selecting sentences with similar intelligibility slopes	145
4.4	Phase III	150
4.4.1	List compilation	150
4.4.2	Experiment I: List application in group of normal-hearing listeners	158
4.4.3	Experiment II: List application in listeners with simulated loss	165
4.4.4	Comparison of results from Experiment I and II	173
4.5	Conclusion	174
4.6	Summary	175
5.	DISCUSSION	176
5.1	Introduction	176
5.2	Phase I: Compiling and refining test materials	176
5.2.1	Compilation of sentences	176
5.2.2	Rating of naturalness	179

5.2.3	Rating of grammatical complexity.....	181
5.2.4	Recording and editing of material.....	182
5.3	Phase II: Selecting an equivalent subset of sentences.....	183
5.4	Phase III: Compilation and evaluation of lists	190
5.4.1	List compilation.....	190
5.5	Conclusion.....	210
5.6	Summary	212
6.	CONCLUSIONS AND RECOMMENDATIONS.....	213
6.1	Introduction.....	213
6.2	Conclusions	213
6.2.1	Main aim: Development of a valid and reliable Afrikaans test for sentence recognition thresholds in noise.....	213
6.2.2	Sub-aim 1 (Phase I): Development of a collection of recorded Afrikaans sentences suitable for the assessment of speech recognition in noise	215
6.2.3	Sub-aim 2 (Phase II): Selecting from the recorded material a collection of sentences with equivalent intelligibility in the presence of noise	215
6.2.4	Sub-aim 3 (Phase III): Comparing inter-list reliability and response variability of two list sets compiled using two different methods of list compilation	216
6.2.5	Research question 1: What methods for the development of a test for sentence recognition in noise have been documented in the literature, and how successful were these methods?	218
6.2.6	Research question 2: Is it possible to improve or streamline previously reported methods that will make the development of such a test more efficient while still producing a reliable measure?	218
6.3	Implications of findings	219
6.4	Critical evaluation of research	220
6.5	Recommendations for further research	225
6.6	Final conclusion.....	227
	REFERENCES	228

LIST OF FIGURES

Figure 2.1: Variables involved in speech audiometry (Adapted from Lyregaard, 1997:35)	28
Figure 2.2: Factors influencing test performance	59
Figure 3.1: Design-based research cycle as applied to the current study	81
Figure 3.2: Research design of the second phase	105
Figure 3.3: Mean intelligibility slope, illustrating the method followed in selecting sentences with a similar slope	112
Figure 3.4: Research process of Phase III	114
Figure 3.5: Intelligibility slopes for lists grouped according to numerical order	116
Figure 4.1: Percentage of sentences derived from each source	135
Figure 4.2: Number of sentences allocated each of the different grammar ratings (1 = simplest; 7 = most complex)	138
Figure 4.3: Distribution of grammar ratings across 515 sentences	139
Figure 4.4: Sentences selected after first equalisation procedure according to intelligibility scores at SNR-5 dB	140
Figure 4.5: Gender differences as demonstrated by results from 1 st equalisation procedure	141
Figure 4.6: Mean percentage intelligibility score as a function of grammar rating	142
Figure 4.7: Mean performance of subjects arranged according to presentation order, showing apparent trend of improvement	144
Figure 4.8: Mean scores at each SNR. Mean differences and standard deviation of these differences (indicated by \pm) are also shown	146
Figure 4.9: Mean scores for male and female participants at SNR-8 and SNR-2	147
Figure 4.10: Intelligibility scores for SNR-2 and SNR-8 as a function of grammar rating	148
Figure 4.11: Mean performance at SNR-2 and SNR-8 arranged according to presentation order. Trendlines indicate practice effect.	149
Figure 4.12: Intelligibility slope of each of the slope lists after exchange of 14 sentences	151
Figure 4.13: Number of errors in phonetic balance per PB list	153
Figure 4.14: Deviation of phoneme counts for all PB lists as a function of the total phoneme count (all phonemes across all sentences; n = 1078) ...	154
Figure 4.15: Number of errors in phonetic balance per slope list	155
Figure 4.16: Deviation of phoneme counts for all slope lists as a function of the total phoneme count (all phonemes across all sentences; n = 1078)	156
Figure 4.17: Intelligibility slopes of PB lists	157
Figure 4.18: Mean SNR-50 across subjects (n=10) for each of the 22 slope lists (unfiltered). Error bars indicate +/- one standard deviation for each list.	160
Figure 4.19: Mean differences between the average of each slope list and the overall mean. Error bars indicate +/- one standard deviation from these means	161

Figure 4.20: Mean SNR-50 across subjects (n=10) for each of the 22 PB lists (unfiltered). Error bars indicate +/- one standard deviation for each list.	163
Figure 4.21: Mean differences between the average of each PB list and the overall mean. Error bars indicate +/- one standard deviation from these means.	163
Figure 4.22: Unfiltered PB lists arranged in order of mean SNR-50 scores .	165
Figure 4.23: Mean SNR-50 across subjects (n=10) for each of the 22 filtered slope lists. Error bars indicate +/- one standard deviation for each list	167
Figure 4.24: Mean differences between the average of each filtered slope list and the overall mean. Error bars indicate +/- one standard deviation from these means.....	168
Figure 4.25: Filtered slope lists arranged in order of mean SNR-50 scores.	169
Figure 4.26: Mean SNR-50 across subjects (n=10) for each of the 22 filtered PB lists. Error bars indicate +/- one standard deviation for each list.	171
Figure 4.27: Mean differences between the average of each filtered PB list and the overall mean. Error bars indicate +/- one standard deviation from these means.....	171
Figure 4.28: Filtered PB lists arranged in order of mean SNR-50 scores	172

LIST OF TABLES

Table 1.1: Summary of dissertation contents by chapter	24
Table 2.1: Advantages and disadvantages of different noise types	40
Table 2.2: Gender of speakers used for different speech-in-noise tests	41
Table 2.3: Selection criteria for speaker as reported in the literature	42
Table 2.4: Results of previous studies on speech-in-noise	47
Table 2.5: Auditory selection criteria of previous studies	50
Table 2.6: Age ranges of subjects in previous studies	52
Table 2.7: Summary of different test method variables	75
Table 3.1: Summary of participant groups	84
Table 3.2: Selection criteria for each group of participants	85
Table 3.3: Rating scale for speakers.....	87
Table 3.4: Description of subjects in Group A.....	90
Table 3.5: Ages of participants in Groups D, E, F, and G	92
Table 3.6: Material and apparatus used in selection of each group of subjects.....	93
Table 3.7: Material and apparatus used during various data collection, recording and analysis procedures	94
Table 3.8: Changes made to sentence content	99
Table 3.9: Composition of playlists for 1 st equalisation procedure	107
Table 3.10: Playlist order for second equalisation procedure	111
Table 3.11: Phonetic symbols used for transcription in spreadsheet	118
Table 3.12: Phoneme occurrences arranged according to frequency of occurrence	119
Table 3.13: Aims and procedures of each phase of the project	132
Table 4.1: Length of sentences derived from each source	135
Table 4.2 Changes made to BKB sentences in translation	136
Table 4.3: Reasons for revision of sentences	137
Table 4.4: Differences in intelligibility scores between grammar ratings	143
Table 4.5: ANOVA results of practice effect for SNR-5 test condition.....	144
Table 4.6: ANOVA results of practice effect for SNR-8 and SNR-2 test conditions	149
Table 4.7: Means and range of intelligibility scores of slope lists at each SNR	151
Table 4.8: Means and range of intelligibility scores of PB lists at each SNR	157
Table 4.9 Means and standard deviations of SNR-50 for each subject across the 22 slope lists (unfiltered).	159
Table 4.10: Means and standard deviations of SNR-50 for each subject across the 22 PB lists (unfiltered).	162
Table 4.11: Comparison of unfiltered PB list pairs	164
Table 4.12: Means and standard deviations of SNR-50 for each subject across the 22 filtered slope lists.....	166
Table 4.13: Comparison of filtered slope list pairs	168
Table 4.14: Means and standard deviations of SNR-50 for each subject across the 22 filtered PB lists.....	170
Table 4.15: Comparison of filtered PB list pairs	172
Table 4.16: Comparison of results from two experiments	173

Table 5.1: Sources used for sentence material by current and previous studies	177
Table 5.2: Characteristics/criteria of sentences used in previous and current research.....	178
Table 5.3: Sentence length reported by previous and current researchers..	178
Table 5.4: Results of naturalness rating of previous and current research ..	180
Table 5.5: Intelligibility slopes of previous and current studies	184
Table 5.6: Percentage of original sentences retained in previous and current studies	186
Table 5.7: Mean SNR at 50% with deviations shown for current and previous studies.....	187
Table 5.8: Difference in list equivalence between two list sets, demonstrated by differences (in percentage) between best and worst scoring lists at each SNR.	191
Table 5.9: Comparison of phonetic balance of current and previous studies.....	192
Table 5.10: List equivalence of current and previous sentence lists	193
Table 5.11: Within-subject score ranges.....	196
Table 5.12: Number of errors in phonetic balance for PB lists, arranged in ascending order	197
Table 5.13: Number of errors in phonetic balance for slope lists, arranged in ascending order	200
Table 5.14: Comparison of results for slope lists, with and without problematic lists	203
Table 5.15: Values calculated separately for Groups 1 and 2* of the slope lists, compared to values for all 22 lists	204
Table 5.16: Comparison of results for PB lists, with and without problematic lists	206
Table 5.17: Comparison of slope lists and PB lists with problematic lists excluded	208
Table 5.18: Comparison of slope lists and PB lists (excluding problematic lists) to existing literature	211
Table 5.19: Different types of validity and reliability and application thereof to the current test.....	212
Table 6.1: Critical evaluation of test method variables as applied in current study	221

LIST OF APPENDICES

Appendix A	Informed Consent Form
Appendix B	Informed Consent Letter
Appendix C	Profile of Participant (Group A)
Appendix D	Rating of Speakers
Appendix E	Case history form (Groups D-G)
Appendix F	Instructions for Rating of Naturalness
Appendix G	Test Form – Slope Lists
Appendix H	Test Form – Phonetically Balanced Lists

LIST OF ABBREVIATIONS

ANOVA	Analysis Of Variance
APD	Auditory Processing Disorders
ARW	Afrikaanse Reseptiewe Woordeskattoets
BKB	Bamford-Kowal-Bench sentences
BKB-SIN	Bamford-Kowal-Bench Speech-In-Noise test
dB	Decibel
dB HL	Decibel Hearing Level
dB SPL	Decibel Sound Pressure Level
HINT	Hearing In Noise Test
Hz	Hertz
kHz	Kilohertz
LTASS	Long Term Average Speech Spectrum
Max	Maximum
Min	Minimum
MTB	Multi-talker Babble (noise)
PB lists	Phonetically Balanced list(s) (specifically referring to the phonetically balanced lists compiled as part of the current research)
PBC	Phonetically Balanced Children's word list
QuickSIN	Quick Speech-In-Noise test
SNR	Signal-to-Noise Ratio
SRT	Speech Reception Threshold
SSN	Speech Spectrum Noise
Std dev	Standard Deviation
WIN	Words-In-Noise test

1. INTRODUCTION AND ORIENTATION

1.1 Introduction

Communication, sometimes called the human connection, is an integral part of our everyday functioning. It is among the most complex of human behaviours and is even considered by some to be the most important human function of all (Bayles and Kaszniak, 1987:47). A fundamental component of this communicative process is the speech signal (Konkle and Rintelmann, 1983:2). Human ears hear best at exactly those frequencies contained in human speech (Northern and Downs, 2002:1), a striking illustration of the fundamental interdependence between hearing and speech.

The importance of our oral communication is what makes us so uniquely human (Northern and Downs, 2002:1,2). The presence of a hearing loss causes a breakdown in this important function of communication and can have a devastating effect on an individual's life. In the paediatric population a hearing loss, if not diagnosed timely, could lead to delays in speech and language development, as well as social, emotional and academic problems (Northern and Downs, 2002:2). Among the elderly, common effects of an untreated hearing loss include sadness and depression; worry and anxiety; paranoia; less social activity; and emotional turmoil and insecurity (National Council on the Aging, 1999:2). These profound effects underscore the importance of accurate assessment of, and intervention for, affected individuals.

The primary problem associated with hearing loss is the reduced ability to hear speech, limiting the individual's communicative effectiveness (Hammond, 1987:1), and the handicap associated with the hearing loss is considered to be equivalent to the ability to understand speech (Barfod, 1979:430). Every day, human listeners use this important communication function of hearing in a great variety of listening situations, most of them characterised by interfering

background noise (Kalikow, Stevens and Elliot, 1977:1338). Background noise has long been known to influence the listener's ability to comprehend speech (White, 1980:158). The ability to understand speech under noisy conditions therefore comprises one of the most important skills for effective communication (Vaillancourt et al., 2005:358).

Due to the importance of this function, the ability to understand speech in the presence of noise constitutes an important area of assessment in audiology. Measuring a patient's ability to discern speech in noise not only quantifies one of the main complaints of individuals with hearing impairment, but also provides the audiologist with valuable information needed for successful rehabilitation of these patients (Smits, Kramer and Houtgast, 2006:538). For this reason, the development of tests to assess this function should enjoy priority within the field of speech audiometry, especially within contexts such as South Africa where so little work has been done in this area.

1.2 Background

Since the human voice is such an essential auditory stimulus in human communication, it has long been regarded an indispensable tool in auditory evaluations. The assessment of pure tone thresholds alone is of limited use in the prediction of communicative deficits, necessitating the use of test stimuli that more realistically represent everyday listening. In this regard, speech stimuli display a great amount of apparent validity (Konkle and Rintelmann, 1983:2). For this reason, the evaluation of a person's ability to hear and understand speech has long been considered an important part of the audiologic test battery. In fact, much of the early work on speech perception was initiated before the professional field of audiology existed (Lucks Mendel and Danhauer, 1997:1). Urbantschitsch (1895 in Silverman, 1983:11) reported that Ernaud in 1761, Pereire in 1767 and Itard in 1805, among others, all experimented with speech training and responses of hearing-impaired children.

Systematic attempts to evaluate speech intelligibility began as early as 1910, when practical methods of assessing telephone channels were developed by Campbell (Lucks Mendel and Danhauer, 1997:1). During World War II, the testing of communication systems by Bell Telephone Laboratories facilitated further developments in the field and is considered by some to be the origin of modern audiology (Silverman, 1983:11). At the same time, researchers at the Harvard Psychoacoustic Laboratory developed a battery of tests for the same purpose of testing communication systems (Egan, 1948 and Hudgins, Hawkins, Karlin, and Stevens, 1947 in Lucks Mendel and Danhauer, 1997:2).

Since these early beginnings, many tests of speech perception have been developed and used for a variety of clinical and research purposes (Lucks Mendel and Danhauer, 1997:2-3). Today, gathering information about a hearing-impaired client's ability to handle speech input is inherent to the basic audiologic test battery (Rupp, 1980:67) and a variety of speech perception tests have been developed with various purposes. Some of the most common purposes of speech perception tests are measuring communication efficiency, determining the degree of handicap caused by a hearing loss, aiding in the selection of amplification, monitoring progress and amplification settings, classifying degree of loss and site of lesion, and serving as a baseline measure for other test procedures (Lucks Mendel and Danhauer, 1997:3; Bess, 1983:127).

Ever since the late 1970's, audiologists have mostly used two basic speech audiometric procedures in routine clinical practice: a threshold test called the "speech reception threshold" and the supra-threshold "word discrimination score" or "speech discrimination test" (Rupp and Stockdell, 1980:16; Wilson and Margolis, 1983:88). The threshold in this case refers to a threshold of intelligibility - defined by Silverman (1983:15) as the point at which the listener understands half of the presented material. This threshold is usually determined through the use of bi-syllabic or spondaic words as test material (Rupp and Stockdell, 1980:16). The supra-threshold word or speech

discrimination test is used to determine the performance level (percentage of test items discerned correctly) at a specific sound intensity that is controlled by the test administrator (Wilson and Margolis, 1983:88).

Besides these two commonly used tests, there are a myriad of other speech perception tests available for auditory assessment, as listed by Lucks Mendel and Danhauer (1997:83-99). Tests can be grouped according to an array of variables, including the type of stimulus materials presented. Different types of stimulus materials include monosyllabic words, spondaic words, nonsense syllables and sentences. Some of the earliest researchers of speech audiometry (Fletcher and Steinberg, 1929 in Silverman, 1983:11) identified two fundamental criteria for speech test material, namely how well it represents everyday speech and how easily the results could be quantified. They noticed from the outset that these two criteria were frequently in conflict. Tests aimed at the analysis of the accuracy of reception of certain fundamental speech units required materials that could easily isolate these units, such as consonants or syllables. These were called articulation tests. In contrast, those tests that used materials more representative of everyday speech, such as sentence materials, were labelled intelligibility tests.

Much of what is generally referred to as “speech audiometry” today, was what Fletcher and his associates called articulation testing (Silverman, 1983:12), in other words tests aimed at isolating and quantifying fundamental units of speech, such as word lists. The spondee threshold test (often called speech reception threshold¹) and the word recognition score (or speech discrimination test) previously discussed are examples of such tests. As mentioned earlier, these tests are quite popular in clinical practice (Rupp and Stockdell, 1980:16; Wilson and Margolis, 1983:88), used by more than 90% of audiologists as a routine testing procedure (Strom, 2003).

¹ Although this test is commonly called a “*speech reception threshold*”, the word “speech” in the term is misleading, as the test really only determines a threshold for spondaic words (Konkle and Rintelmann, 1983:6).

Despite its common usage, these tests show various limitations. Both of these tests use single words as test stimuli, which shows a very poor resemblance to everyday listening situations (Lucks Mendel and Danhauer, 1997:62; Konkle and Rintelmann, 1983:5). Additionally, monosyllabic materials demonstrate limited efficiency to differentiate among amplification options (Bess, 1983:188). Furthermore, presenting these materials in quiet when common listening situations almost always involve some degree of background noise has been criticised by theoreticians and patients alike (Carhart, 1965 in Lucks Mendel and Danhauer, 1997:62; Killion, 2002:63).

The validity of test materials in terms of its representation of everyday speech signals, also called “ecological validity” (Mackersie, 2002:395), appears to be greater in the case of sentence material (Bess, 1983:140; Konkle and Rintelmann, 1983:5). By integrating all of the aspects of auditory ability into a single performance measure, sentence tests tend to give a more global impression of an individual’s speech perception than single words (Lutman, 1997:82). Moreover, tests using sentences as stimuli have demonstrated increased sensitivity to changes in performance as a function of intensity changes (which can be called the intelligibility slope), especially when used in the presence of noise. Plomp and Mimpen (1979:49) and Kollmeier and Wesselkamp (1997:2415) both reported unusually steep slopes of intelligibility with the use of sentence materials in noise to determine speech recognition thresholds. This means that small changes in signal-to-noise ratio (SNR) yielded notable changes in speech recognition, increasing the precision with which small differences in speech recognition can be detected.

However, the use of sentence material in speech audiometry is not without its limitations. A frequently raised concern is the redundancy of the material, which makes the test too easy, as some sentences may be correctly identified by recognising a single word (Owens, 1983:359). The context provided in sentence materials contributes heavily to intelligibility, which could lead to better performance. However, it still stands that this is representative of typical

everyday functioning, where verbal communication is often loaded with redundant cues. Furthermore, speech perception tests using sentence material can be made more difficult by adding a competing signal, such as a background noise (Bess, 1983:168). Another argument against sentence material is that once a sentence is known or familiar to the subject, even fewer cues are sufficient for recognition (Owens, 1983:359). However, this merely implies that tests using sentence material should consist of a large collection of sentences to avoid familiarisation to the material (Van Wieringen and Wouters, 2006:3). These sentences could then be grouped into lists that can each be used for testing at separate occasions.

The arrangement of sentences into lists poses another challenge to developers of such tests, namely the equivalence of the different test lists (Fletcher and Steinberg, 1929 in Silverman, 1983:13). This issue, however, is not unique to sentence materials, as a great deal of effort has been spent on creating equivalent single-word lists for speech audiometry (Egan, 1948 in Bess, 1983:166). Despite these efforts, the use of only half of the items in a monosyllabic word list is a common in clinical practice (Bess, 1983:164) which inevitably compromises the phonetic balance, and therefore possibly the equivalence in difficulty between lists (Roets, 2005:67).

Sentence lists consist of more complex stimuli than a monosyllabic word test, which means that there are more variables affecting the equivalence of test items than there is with word lists. However, multiple researchers have managed to compile reasonably equivalent lists by phonetically balancing their sentence material (Plomp and Mimpen, 1979:45; Nilsson, Soli and Sullivan, 1994:1088; Hällgren, Larsby and Arlinger, 2006:229; Vaillancourt et al., 2005:362; Wong and Soli, 2005:282). Therefore, despite the complexity of the material, the feasibility of compiling a sentence test with equivalent lists has been repeatedly demonstrated.

Speech audiometric tests using sentence material as stimuli therefore comprise a viable and important aspect of hearing assessment in both experimental and clinical settings, showing a high degree of ecological validity. However, the listening condition in which speech tests are used also affects the ecological validity of these tests. It has long been recognised that testing speech perception in an adverse listening condition will give a much better indication of the individual's functioning in normal life than testing in quiet (Carhart, 1968:715). The ability to understand the complex speech signal depends not only on peripheral hearing sensitivity, but also on other auditory skills such as frequency resolution, temporal resolution, suppression and intensity discrimination, which collectively is known as distortion (Lutman, 1997:63). Evaluating merely the peripheral hearing sensitivity with a test applied in an ideal listening environment will thus give an incomplete impression of a patient's auditory performance. This fact becomes evident when speech perception is evaluated under noisy conditions, as the direct effect of sensitivity is removed and the correlation between intelligibility and sensitivity becomes an indirect relationship mediated through distortion (Lutman, 1997:65).

The redundancy of the speech signal in favourable listening conditions enables listeners to follow a conversation with ease. In fact, since the information needed for understanding is spread so redundantly over wide parts of the frequency range, even narrow bands of information are sufficient to facilitate understanding (Barfod, 1979:433). However, as listening conditions deteriorate, those with normal hearing may still be able to understand speech, while its intelligibility will break down for individuals who suffer from a hearing loss (Versfeld, Daalder, Festen and Houtgast, 2000:1671). The hearing loss robs the speech signal of some of its redundancy and makes it more vulnerable to the effect of noise (Barfod, 1979:433). This may be why the most common complaint of patients with sensorineural hearing loss is difficulty to understand speech in situations with background noise (Smits et al., 2006:538).

Since the ultimate intelligibility of a speech signal depends not only on whether it is audible for the listener, but also on the degree to which the auditory system can make use of the signal (Gatehouse and Robinson, 1997:79), it is not possible to predict a patient's ability to understand speech in noise from the results of their pure tone audiogram (Killion and Niquette, 2000:50). For this reason, the measurement of a "signal-to-noise ratio loss" or SNR loss has been suggested (Killion, 2002:59). This is defined as the increase in SNR (when compared to the SNR that normal-hearing subjects need) required by an individual to enable 50% correct repetition of words in a sentence. Measuring this loss will enable audiologists to assess and quantify that which appears to be most important to patients with hearing impairment, namely their ability to understand speech in noise (Killion, 2002:60). It could assist audiologists in determining why two patients with comparable hearing losses and identical hearing aids may be having vastly different experiences in terms of hearing aid benefit (Killion, 2002:60). This could result in more effective patient counselling by providing valuable prognostic and rehabilitative guidelines.

1.3 Rationale

Since speech audiometry utilising sentence materials, especially when used for the evaluation of speech perception in noise, is an invaluable tool in any audiology practice, it is unfortunate that there seems to be no such standardised measures in the South African context. A study by Roets (2005) investigated the current practices of South African audiologists in terms of speech audiometry and compared this to ideal practice. The findings of this study indicated several weaknesses in South African practices.

In terms of test content, the study revealed a great paucity of standardised material in many of the official languages (Roets, 2005:130). The ideal criteria for stimulus material in speech perception tests stipulate the inclusion of a combination of single words, nonsense syllables and sentences (Roets, 2005:84). However, only one of the 84 respondents in the South African study

reported using sentence material in speech audiometry procedures (Roets, 2005:75). The test reportedly used by this respondent was the *Synthetic Sentence Identification* test (Speaks and Jerger, 1965), which was not developed for the South African population. Furthermore, none of the other participants were able to name any standardised sentence tests for speech audiometry. These findings elucidate the need for the development of standardised South African sentence material, along with an increased awareness among practitioners of the importance of using such materials in speech audiometry.

A further problem noted within the South African context is the lack of pre-recorded stimulus material. Although one respondent in the South African study reported knowledge of a locally recorded version of English and Afrikaans stimulus material, not a single respondent reported using this (Roets, 2005:100). This is despite the literature consensus that monitored live voice presentation methods substantially reduce test reliability (Konkle and Rintelmann, 1983:7). Therefore, the standardisation of pre-recorded materials and promotion of its clinical usage are neglected priorities that need to be addressed.

As far as the assessment of speech perception in noise is concerned, locally standardised tests also seem to be in short supply. Unfortunately the South African survey by Roets (2005) did not investigate speech-in-noise tests and surveys of a similar nature are not mentioned in the literature. A general practice appears to be the application of tests standardised in quiet with the speech noise on the audiometer used as a background noise. Ostergard (1983:233) cautions against this practice, as this would violate the assumption of standardisation that test administrators would maintain the same conditions used during standardisation during application of the test when comparing results to set norms. This argument is supported by the findings of Stockley and Green (2000:91), who applied the Northwestern University Auditory Test No. 6 (NU-6) lists to both normal-hearing and hearing-impaired listeners in

quiet and in noise. Their findings revealed that these lists were equivalent in difficulty when applied in a quiet condition to both groups of listeners, but not in the presence of background noise.

This indicates that the reliable assessment of speech perception in noise requires the development and standardisation of test material designed specifically with this purpose in mind. The value of a test of speech perception in noise lies partly in the extent to which it represents everyday listening. Therefore, sentences are a judicious choice of material for such a test, partly due the redundancy of the content (Owens, 1983:359), and partly since sentences closely resemble everyday speech signals, thereby serving to further enhance the ecological validity of the speech-in-noise test.

1.4 Problem statement

It is clear that there is a need for the development of tests in South African languages using sentence material for the assessment of speech perception. This material should be pre-recorded and should provide a means to assess and quantify the ability of a listener to understand speech in the presence of background noise. Furthermore, the format of the test should facilitate fast and easy administration, as this will enhance the clinical usefulness of the test in busy clinical practices. Such tests will improve service delivery to individuals with hearing impairment by providing a more reliable testing procedure than monitored live voice procedures; by utilising stimulus materials that are more representative of daily input; and by enabling audiologists to evaluate and quantify that which constitutes one of the most common complaints of hearing aid users, namely understanding speech in the presence of background noise.

However, the linguistic diversity of the South African population makes the development of such tests a complicated issue. In this multi-cultural context there are eleven official languages (Department of Arts and Culture, 2002:5),

and in many of these languages, no standardised speech audiometry tests have been developed or published, emphasising the need for the development of such tests. The immense task of developing standardised speech audiometry material in all of the official languages should be approached as any vast enterprise – one step at a time. The study at hand will thus focus on one of the official languages and will serve a dual purpose – the development of a useful tool for measuring speech recognition in noise in that specific language, and the establishment of an efficient template method for developing similar tests in all the other languages.

The language selected for the present research is Afrikaans. According to Statistics South Africa (2001:5), Afrikaans is the third most common home language in the country, spoken by 13.3% of the population. In Gauteng, the province where the current research was conducted, Afrikaans is the second most common home language, reported by 14.4% of the population to be their first language (Statistics South Africa, 2004:21). Since there are no standardised or published tests of sentence recognition in noise in Afrikaans, the compilation of such a test comprises a valuable asset towards an ideal test battery that consists of different types of material for speech audiometry for the diverse South African context (Roets, 2005:65).

From the preceding discussion, the following questions emerge.

- ❑ What methods for the development of a test of sentence recognition in noise have been documented in the literature, and how successful were these methods?
- ❑ Is it possible to improve or streamline previously reported methods that will make the development of such a test more efficient while still producing a reliable measure?

The main aim of the current study will therefore be ***the development of a valid and reliable Afrikaans test of sentence recognition in noise***. In the

process of developing this test the questions listed above will be addressed by:

- ❑ critically reviewing existing literature to explore previously reported methods;
- ❑ developing a suitable methodology for the development of an Afrikaans test of sentence recognition in noise; and
- ❑ employing and assessing the reliability of novel methods that may be more efficient than previous methods in the development of this test.

The specific sub-aims formulated to address these questions will be specified and described in Chapter 3.

1.5 Definition of terms

- ❑ ***Speech audiometry:*** The term “audiometry” literally refers to the measurement of hearing (Konkle and Rintelmann, 1983:1). “Speech audiometry” in turn refers to the variety of techniques that use speech stimuli to assess auditory function. Throughout the present study, it will refer to different tests, including those presenting speech in the presence of background noise.
- ❑ ***Speech perception test:*** Throughout the study, this term will refer to a test aimed at providing a measure of how well listeners comprehend speech (Lucks Mendel and Danhauer, 1997:3). For the purposes of the study at hand, this will also include tests that evaluate the understanding of speech in the presence of background noise. Tests designed for the paediatric population will not be included in discussions within the present study.

- ❑ **Speech recognition:** This term refers to the task of the listener in a speech audiometric procedure and may be used to refer to both threshold and supra-threshold speech perception tests, traditionally referred to as “speech reception” and “speech discrimination” tests respectively (Konkle and Rintelmann, 1983:6; Wilson and Margolis, 1983:89). Within the context of the present study, this term was chosen to describe the type of test that will be developed, since this test is a type of threshold test, but does not resemble the traditional threshold or reception test. The term has also been used in the literature when referring to the type of test that will be developed in this study (Hällgren et al., 2006).

- ❑ **Speech-in-noise test:** This term will be used generically throughout the study to refer to any speech perception test that is routinely presented in noise. This includes a great variety of tests and is not specific in terms of the speech or noise stimuli used in these tests. It could therefore, for example, refer to both a test using single words and speech noise on the audiometer and a test using sentence material and a specifically standardised noise.

- ❑ **Signal-to-noise ratio:** Signal-to-noise ratio (abbreviated SNR) refers to the relationship between the intensity of the speech signal and the intensity of the background noise. An SNR of -5 dB would then, for example, indicate that the speech signal is presented at an intensity of 5 dB less than the noise. “SNR loss” is defined as “*the increased SNR compared to that normally required by a subject to repeat 50% of words in sentences correctly*” (Killion, 2002:59). SNR-50 refers to the SNR required by a subject to correctly discriminate 50% of the presented speech signal.

- ❑ **Intelligibility slope:** Within the present study, this term refers to a graph in which the percentage of items in a speech test that are correctly identified is depicted as a function of the SNR at which these stimuli were

presented. The term is then generally used to indicate the extent to which a specific sentence increases in difficulty as the SNR becomes poorer. For example, a sentence that remains equally easy/hard to understand regardless of changes in the SNR is considered to have a flat intelligibility slope. On the other hand, a sentence that becomes significantly more difficult as the SNR worsens is considered to have a steep slope.

- ❑ **Reliability:** This term refers to the precision of a test in terms of the consistency of test data across multiple similar observations (Ostergard, 1983:224). There are different types of reliability, including coefficient of stability, coefficient of equivalence, and internal consistency (Ostergard, 1983:225). These specific types of reliability will be discussed in greater detail in the second chapter.

- ❑ **Validity:** Essentially, the validity of a given test denotes whether that test measures what it is supposed to measure (Lucks Mendel and Danhauer, 1997:8). Therefore, the validity of a test is not inherent to the test itself, but rather related to its purpose – whether it is capable of achieving its aims (Ostergard, 1983:223). There are different types of validity, and these will be discussed further in the following chapter. It should be noted here, however, that the term “ecological validity”, which is not a traditional type of validity, indicates the extent to which the results of a test “align with the realities of speech perception in natural listening environments” (Mackersie, 2002:395).

- ❑ **Native speaker(s):** Any individual who considers a specific language to be their first language or mother tongue is considered a native speaker of that language. This can be more specifically defined as the language most often spoken at home by the individual (Statistics South Africa, 2001:5). Throughout the study, native speakers will refer to individuals who consider Afrikaans to be the language they mostly speak at home, unless otherwise indicated.

- **SNR-50:** The signal-to-noise ratio at which a listener would be able to perceive 50% of a presented sentence. This is sometimes referred to as “SRT” or Speech Reception Threshold (Hernvig and Olsen, 2005:510), but to avoid confusion with the traditional meaning of SRT, namely the results of a list of spondaic words presented in quiet, the term SNR-50 will be used in this study (Killion and Niquette, 2000:48).

- **Mean or average:** The term “mean” is used throughout the report to refer to the arithmetic average value (Maxwell and Satake, 2006:289). The term “average” is occasionally used to convey the same meaning, mostly in cases where the word “mean” would be confusing (for example when talking about the mean or average of a number of means).

1.6 Outline of chapter contents

The current dissertation provides a detailed description of the procedures followed to address the research questions and needs as described in this chapter. Table 1.1 provides an outline of the content of each of the chapters in the dissertation. This table serves to provide a guide to the structure of the dissertation, as well as a concise summary of the content.

Table 1.1: Summary of dissertation contents by chapter

Chapter 1	<i>Introduction and Orientation:</i> This chapter provided a rationale for the necessity of the study at hand, as well as a conclusive problem statement. By looking at the significance of speech audiometry and deficits in current practice, it provided a motivation for the development of an Afrikaans test for the assessment of sentence recognition in noise. It also supplied clarification of terminology used throughout the study.
Chapter 2	<i>Methodological considerations in the development of a speech-in-noise test:</i> The second chapter explores the aspects involved in the development of the test discussed in the first chapter. A review of existing literature on the subject is provided according to the variables involved in speech audiometry. The operationalisation of these aspects provides the foundation for the following chapter.
Chapter 3	<i>Methodology:</i> The third chapter describes the methodology of the study by specifying the aims, research design, ethical considerations, research sample, material and apparatus, procedures, as well as the validity and reliability of the research.

- Chapter 4** *Results:* Results obtained during each of the three research phases are depicted in this chapter.
- Chapter 5** *Discussion:* This chapter provides an interpretation of the results described in the previous chapter. The variables involved in the developed test are discussed and reviewed in light of previous studies found in the literature.
- Chapter 8** *Conclusions and Recommendations:* This final chapter draws conclusions on the findings of the present study in referral to the research aims and the problem statement of the first chapter. A critical evaluation of the developed test as well as the methodology followed in its development is provided along with a discussion of the contributions that the study makes. Recommendations for further research are made in view of the current findings, suggesting future applications of and improvements on the developed measure.
-

1.7 Conclusion

Speech audiometric procedures comprise an essential part of the audiologic test battery. Unfortunately, speech perception tests that are commonly used in clinical practice show several deficiencies and limitations. The use of single words as test stimuli, as well as the presentation of tests in unnaturally quiet conditions limits the extent to which test results can be generalised to predict everyday functioning. In South Africa, the paucity of standardised tests and pre-recorded material further depreciates the accuracy and reliability of speech audiometry. All of these factors underscore the need for the development of a local test to measure speech perception in noise, preferably using ecologically valid stimulus material, such as sentences.

1.8 Summary

This chapter provided an overview of the importance of speech audiometry in the evaluation of hearing-impaired individuals. It also provided a motivation for the use of sentences as stimulus material in speech perception testing. Additionally, the value of testing speech perception in noise was elucidated, and the current deficits in these test methods, especially within the South African context, were indicated. Several keywords that will be used throughout this report were clarified, and an outline of the chapter contents was provided.

2. METHODOLOGICAL CONSIDERATIONS IN THE DEVELOPMENT OF A SPEECH-IN-NOISE TEST

2.1 Introduction

The inability to understand speech in the presence of background noise is considered by many audiologists to be the most common complaint of adult patients with a hearing loss (Wilson and McArdle, 2005:81). Measuring a patient's ability to discern speech in noise therefore not only quantifies one of the main complaints of individuals with hearing impairment, but also provides the audiologist with valuable information needed for successful rehabilitation of these patients (Smits et al., 2006:538). In addition, the findings of a speech-in-noise evaluation are particularly valuable when counselling patients with regards to their expectations of amplification systems (Wilson and McArdle, 2005:82). For these reasons, the development of tests to assess this function should enjoy priority within the field of speech audiometry, especially within contexts such as South Africa where so little work has been done in this area.

However, as the perception of speech is a complex process affected by a large number of factors (Lyregaard, 1997:26), any test measuring this ability will also be influenced by a myriad of variables. When developing a new test of speech perception, these variables must receive careful consideration (Lyregaard, 1997:37). Since the value of sentence material in speech audiometry and the rationale for developing a speech-in-noise test using sentences has already been established in the first chapter, all discussions will henceforth focus mainly on tests of sentence recognition in noise, as opposed to other types of speech perception tests.

2.2 Framework for the development of speech-in-noise tests

To provide a theoretical background for the development of a speech-in-noise test, it is necessary to explore the variables involved in the process of speech audiometry by critically analysing the literature on the subject. In order to structure the large collection of variables involved in this process, it is necessary to organise the variables into a framework of speech audiometry. This will provide researchers with a model that encompasses all the important variables within the process of speech audiometry.

Lyregaard (1997:35) provided such a model in the form of a block diagram illustrating the major variables in speech audiometry. This diagram has been adapted for the present discussion in order to clearly illustrate all the major variables involved. Figure 2.1 provides an illustration of the adapted model. Although this illustration is by no means exhaustive, it provides a framework or outline for a more in-depth discussion of each of these components. Each of these components will subsequently be discussed in the light of existing literature. It should be noted here that the purpose of the discussion is to clarify these variables as they relate to the development of a new test, and not to provide an exhaustive theoretical model of speech perception.

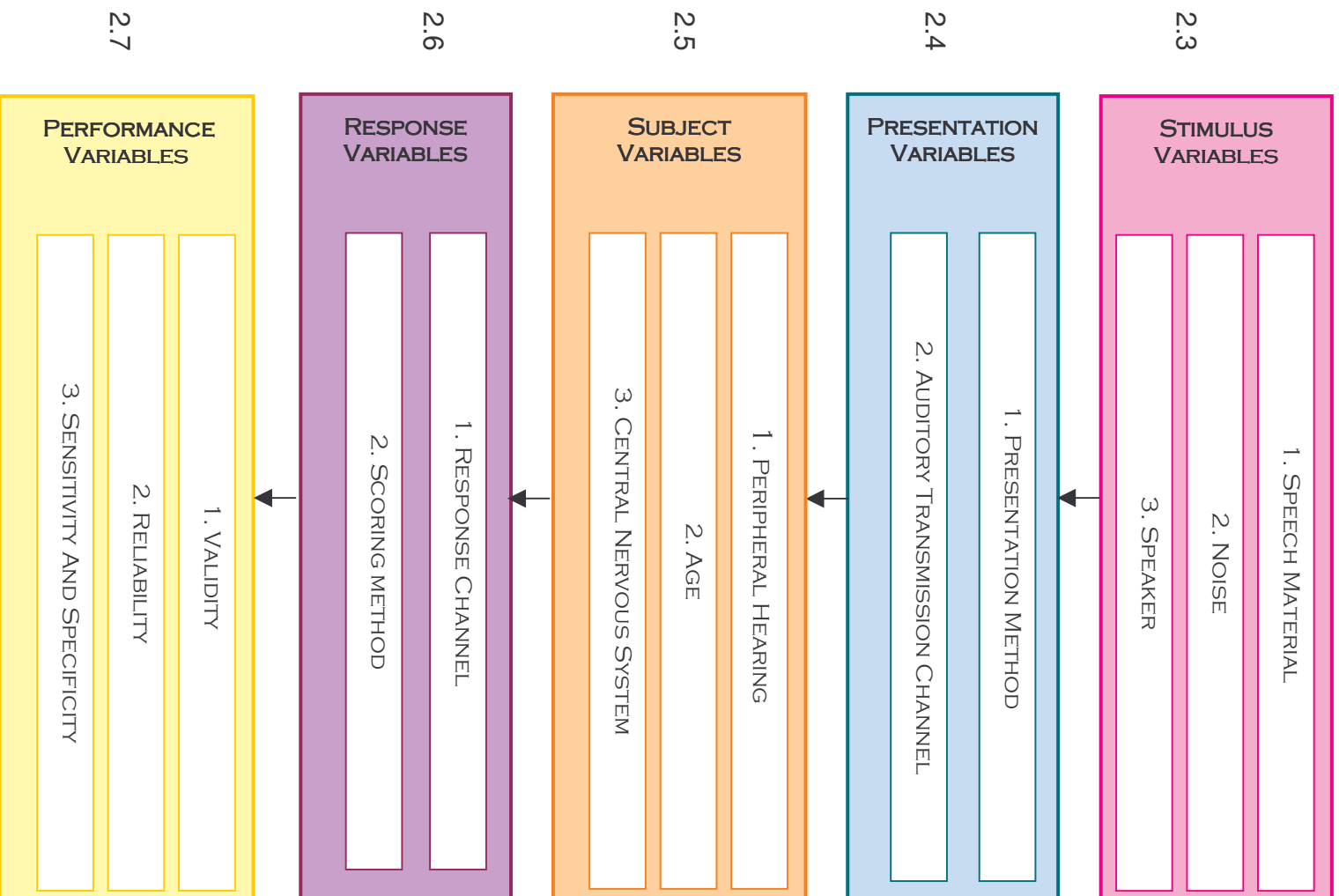


Figure 2.1 : Variables involved in speech audiometry (Adapted from Lyregard, 1997:35)

The following sections will provide a critical discussion of each of these variables within the context of available literature. At the end of the chapter, the options considered under each variable are summarised in table format.

2.3 Stimulus variables for speech-in-noise tests

As illustrated in Figure 2.1, the stimulus variables in speech audiometry include the speech material, the noise and the speaker. This section will provide an overview of stimulus materials developed or used by previous researchers. The purpose of the discussion will be to critically review different methodologies used in the development of stimulus material in order to find a method that is both efficient and suitable for the development of such a test in Afrikaans.

2.3.1 Speech material

Since 1947 researchers worldwide have developed a great variety of speech perception tests using sentence material as stimuli (Lucks Mendel and Danhauer, 1997:73). Sentence materials vary from synthetic (nonsense) sentences to regular, predictable sentences to multiple-choice options, depending on the purpose of each test. Popular speech tests using sentence material developed in the past include the Central Institute for the Deaf (CID) Everyday Sentences (Silverman and Hirsch, 1955), Synthetic Sentence Identification (Speaks and Jerger, 1965), Speech Perception in Noise Test (Kalikow, et al., 1977) and the Hearing In Noise Test or HINT (Nilsson et al., 1994) among others.

2.3.1.1 Sources of speech material

Past composers of test material for a speech-in-noise test have followed mainly one of two methods. Some researchers have developed their own, original material, whereas others have translated and/or adapted material from other tests. Both of these methods have been used successfully and

have occasionally been combined, with equally effective results, as seen in the discussion below.

One of the earliest reports on a test using meaningful sentence materials with added noise is that of Plomp and Mimpen (1979). These researchers created an original set of sentences, keeping to certain criteria. The criteria specified that sentences had to represent conversational speech; be short enough to repeat; exclude proverbs, questions and exclamations; and consist of 8 or 9 syllables (Plomp and Mimpen, 1979:44). The experiments conducted with the material proved that the sentence collection could be used for reliable and sensitive measurements of speech recognition (Plomp and Mimpen, 1979:49).

Contrary to this, the developers of the American Hearing in Noise Test (HINT) (Nilsson et al., 1994) did not create an original set of sentences, but rather used the sentence material from the Bamford-Kowal-Bench or BKB sentences (Bench and Bamford, 1979). The BKB sentences are a large set of short sentences based on the expressive language of hearing-impaired children in the age range 8-15 years (Blandy and Lutman, 2005:436). These sentences were selected by the developers of the HINT due to the size of the collection as well as its simplicity. The content was adapted for use in the United States by removing British idioms, equating sentence length and subjecting it to an evaluation of naturalness by native speakers of American English (Nilsson et al., 1994:1086).

Other researchers have also followed this trend of adapting test material. According to Wong and Soli (2005:276), who adapted the HINT for Cantonese, the test has also been adapted in Japanese, Latin American and Spanish. Hällgren et al. (2006) compiled a Swedish version of the HINT. These authors used the HINT sentence material as a starting point, and adapted it to the Swedish language. Sentence content was adjusted after having the material rated for naturalness by native speakers. Kollmeier and Wesselkamp (1997) developed a German sentence test for intelligibility

assessment by combining test materials from existing German sentence tests and editing these before making a new digital recording of the material (Kollmeier and Wesselkamp, 1997:2412).

In Canada, Vaillancourt and associates (2005) adapted the HINT for French-speaking Canadians. However, the authors did not use the HINT or BKB sentences, but developed original sentence material (520 sentences in total), which was derived from a corpus of nouns and verbs associated with a grade-one (6-7 year old children) comprehension level. These sentences were revised to eliminate idioms, proper nouns and repetitive phonemes and were also rated for naturalness by native speakers. Likewise, the developers of the “Listening in spatialized noise-sentences” (LISN-S) (Cameron and Dillon, 2007a:199) had a new set of sentences developed by speech pathologists specialising in the rehabilitation of children with hearing impairments. These sentences were formulated according to the criteria used in the development of the BKB sentences.

In South Africa, Olivier (2000) also followed the method of translating and adapting existing test material. This author used the BKB sentences to develop a speech perception test in Xhosa. After translation, sentence material was evaluated and sentences with British terminology, syntactical, and/or semantical errors were eliminated. The recorded sentences were then presented to a group of Xhosa-speaking individuals with normal hearing to test the suitability of the vocabulary in terms of dialect, concepts and age group. Afterwards, the material was arranged into phonetically equivalent lists and a recorded version thereof was presented to a second group of listeners to evaluate list equivalency (Olivier, 2000:66).

Versfeld et al. (2000:1672) followed a somewhat different approach in their compilation of test material, since their aim was to create an unusually large set of Dutch sentences that could yield better test efficiency than existing Dutch tests. These authors used an automated selection process to select a

set of 35 000 sentences from large databases of sentences that were available in digitised format. Following this, sentences underwent a manual selection process to ensure that it complied with the criteria stipulated by Plomp and Mimpen (1979:44). The 1500 selected sentences were then reviewed by a group of experts in the field of speech therapy and audiology who made additional recommendations for adjustments or exclusion of sentences, yielding a remaining set of 1311 sentences.

The Dutch sentences developed by Versfeld et al. (2000) proved to be too difficult for individuals with a severe hearing impairment and those with cochlear implants. Therefore, Van Wieringen and Wouters (2006) compiled two tests specifically for the quantification of speech understanding in severely impaired listeners. One test featured numbers as the presented material, and the other sentences. These authors developed their own material, and used the selection criteria stipulated by Plomp and Mimpen (1979:44) in the selection of sentences. Results indicated that the developed materials were indeed suitable for the evaluation of individuals with severe hearing losses (Van Wieringen and Wouters, 2006:14).

Wong, Soli, Liu, Han and Huang (2007:71S) created two Mandarin versions of the HINT – one for listeners from mainland China and one for Mandarin listeners from Taiwan. These authors compiled their own sentence material for this goal. Sentences were created by two audiologists – one being a native Mandarin speaker from mainland China and the other from Taiwan. The sentences were equal in length (10 characters each) and were easily understood by people from different educational backgrounds and children aged 6 and up. In view of the fact that both the adaptation of test materials and the creation of new sentences are able to yield reliable test material, it should be possible to combine these two methods in cases where a large body of sentences of sentences is required.

Regardless of the sources used for sentence compilation (translated or originally developed), many of the researchers mentioned here have used at least some of the criteria initially stipulated by Plomp and Mimpen (1979:44) as a guide for the style and content of the sentence material. Most other researchers have developed material that is representative of everyday or conversational speech (Versfeld et al., 2000:1672; Wong and Soli, 2005:278; Hällgren et al., 2006:228; Van Wieringen and Wouters, 2006:4; Wong et al., 2007:71S). A number of studies have stipulated, in accordance with the criteria of Plomp and Mimpen (1979:44), that the sentences should not contain proverbs, exclamations, questions, or proper nouns (Versfeld et al., 2000:1672; Vaillancourt et al., 2005:360; Van Wieringen and Wouters, 2006:4). Some studies have added to the criteria of Plomp and Mimpen (1979:44) that sentences should be syntactically complete, or at least contain a verb and a noun (Kollmeier and Wesselkamp 1997:2413; Versfeld et al., 2000:1672; Van Wieringen and Wouters, 2006:4).

2.3.1.2 Assessing naturalness of speech material

To ensure that the style and content of sentence materials are representative of everyday speech, a number of researchers have had the material rated for naturalness by native speakers (Nilsson et al., 1994:1086; Vaillancourt et al., 2005:360; Hällgren et al., 2006:228; Wong and Soli, 2000:278). This is done to ensure that translated and/or newly developed material is considered acceptable by the general population. Rating was typically done on a scale of 1 (artificial) to 7 (natural), and any sentences receiving a mean rating lower than six was revised and submitted to a second round of rating. The number of subjects asked to rate the sentences varied from four to fifteen per round. The common selection criterion between all the studies was that subjects had to be native speakers of the language in which the test was being developed.

In addition to this common criterion, Vaillancourt et al. (2005:360), although only using a small number of subjects (nine in total), reported that their subjects were purposefully selected to be from different geographical origins,

age groups and educational backgrounds. This was done in an attempt to ensure that the material would be considered natural by all members of the French-Canadian population (Vaillancourt et al., 2005:360). Although this method appears to be a good way of making material acceptable to people spread out over a large geographic area, results should be generalised with caution. The only way to ensure that material is appropriate for people living in a certain area would be to apply it to a representative sample of its population, and not merely a single person. This method does however provide a preliminary indication towards the ecological validity of the material for different groups of people, in a manner that is viable for the purposes of a time-constrained research project.

An additional consideration should be made when submitting material to native speakers from different regions for a rating of naturalness, namely the wide range of varieties of a language in terms of content and pronunciation. Native speakers from different regions of a country may have been exposed to different varieties of the language, including standard and non-standard varieties. A standard variety is defined as the form of a language that serves as a model against which users of the language measure the way they talk or the variety accepted by the language community as the language form with the greatest value for use across the entire language territory (Carstens, 2003:283, 285).

In order to prevent any one specific subject who may have consistently been exposed to a non-standard language variety from exerting too strong an influence on the results, no alterations should be made to sentences that received a low rating from only one of the subjects. A minimum of two subjects should provide a low rating and recommendation for change before a sentence may be altered. Additionally, the pronunciation of the material during recording should aim to resemble a standard pronunciation by avoiding the abnormal and striving for the general form as far as possible (Le Roux and Pienaar, 1976:xviii).

2.3.1.3 Ensuring homogeneity of sentence material

To determine the homogeneity of developed sentence materials, an efficient manner of assessing the difficulty of each sentence must be devised. Nilsson et al. (1994:1090) checked the grammatical and syntactic level of the HINT sentences using a commercial software package that graded the reading level of a text version of the material. Although grading the reading level of material was deemed a simple, repeatable and objective technique of difficulty estimation, some reservations towards this approach should be expressed. First of all, this method is based on the assumption that a word or sentence that is difficult to read is also difficult to understand and vice versa. However, sentences that are judged to be on a similar reading level may not be on the same verbal language level. Secondly, although software to grade readability may be readily available in English, this is not the case for many other languages. In view of these reservations, an alternative approach may be to grade the difficulty of sentences in terms of grammar and syntax, instead of readability.

After the development of a collection of sentences that has been rated for naturalness by native speakers, these materials are usually recorded and then submitted to a procedure for selecting sentences that are equally difficult to understand in the presence of a specific level of noise. This procedure, as first reported by Plomp and Mimpen (1979:44) entails presenting the recorded sentences to a group of normal-hearing listeners under a specified noise condition. The mean performance of the listeners on each sentence is used to determine the chance of correct recognition for each sentence. Following this, researchers may use two different methods to ensure that the sentences in the final collection are equally difficult. Firstly, the average intensity of a sentence may be adjusted (increased if the sentence was too difficult and decreased if the sentence was too easy). This method was followed by Plomp and Mimpen (1979:44); Nilsson et al. (1994:1088); Kollmeier and Wesselkamp (1997:2414); Wong and Soli (2005:279); Hällgren et al. (2006:229) and Wong et al. (2007:71S).

An alternative method is to exclude or eliminate from the collection those sentences that are too difficult or too easy. Plomp and Mimpen (1979:44) combined this method with the re-scaling method described above. Versfeld et al. (2000:1673), Vaillancourt et al. (2005:361), and Van Wieringen and Wouters (2006:4) used this method to equate the difficulty of their sentences. With this method, sentence material is typically presented to normal-hearing subjects at a pre-determined SNR level and a mean percent intelligibility score is determined for each sentence based on the number of words repeated correctly (Vaillancourt et al., 2005:361). Subsequently, a mean percentage score for all the sentences is determined and sentences that do not fall within a determined range of this mean are eliminated. The advantage of this method is that it does not require recruiting and re-evaluating a new selection of subjects for the assessment of the adjusted material, and is therefore more time-efficient. On the other hand, it implies that the initial collection of sentences should be much larger than the final number of sentences required, allowing researchers to eliminate sentences.

However, caution should be applied when determining the homogeneity of a collection of sentences merely by looking at the performance of subjects at one specific SNR. This is because this specific SNR is simply a point on a psychometric function, where psychometric function indicates a graph relating a measure of performance to a stimulus dimension (Wilson and Margolis, 1983:80), such as the intelligibility slope. Therefore, the mere comparison of the percentage scores of two sentences at a fixed SNR, does not take into account the slopes of the psychometric function (performance as a function of SNR), in this case also called the intelligibility slope. This could give an inaccurate reflection of the change in threshold, as the actual slope and/or the maximum score could have changed as well. In other words, just because two sentences yield the same performance by listeners at one SNR, does not imply that it will do the same at another SNR, as this depends on the slope of the intelligibility function. Therefore, researchers must also ensure that the collection of sentences selected have similar intelligibility slopes. Thereby, the

performance of a listener will increase or decrease an equal amount for each of these sentences with variation of the SNR.

It is for this reason that Versfeld et al. (2000:1676) evaluated the psychometric function of each individual sentence in their collection by presenting it to subjects at two different SNRs. This enabled them to select from the collection sentences that were homogeneous in terms of threshold (SNR-50) and slope (Versfeld et al., 2000:1678). This method was similar to that of Kollmeier and Wesselkamp (1997:2413) and Vaillancourt et al. (2005:361) who presented their material to subjects at three different SNRs in order to obtain an intelligibility slope for each sentence. When using this method, sentences are selected on the basis of both their SNR-50 threshold (SNR where 50% correct recognition occurs) and their intelligibility slope.

To reduce experimentation time, it should be possible to reduce the size of the sentence collection first by presenting all the sentences at an SNR where 50% performance is expected and excluding outlying sentences on the basis of this data (Plomp and Mimpen, 1979:44; Hällgren et al., 2006:229; Wong and Soli, 2005:279). Subsequently, the remaining sentences (selected subset) could be presented at two other SNRs in order to establish the psychometric slopes. This will reduce the time needed to test each subject, as the latter two SNR experiments will use a smaller number of sentences. The final set of sentences could then be selected according to both the psychometric slope and SNR-50 of the selected subset.

2.3.2 Type of noise

The early use of noise in speech audiometry used to be limited by the types of noise that could be generated. However, technological developments have brought about the generation of several different types of noise. Since a complex noise was found to be ineffective in the masking of a speech signal, speech noise was developed to mask the narrow band of frequencies surrounding the speech spectrum (Stockdell, 1980:110).

Within the context of speech-in-noise tests, there are different types of noise reported to be efficient maskers of the speech signal. The SPIN test, which is somewhat different in test format to the speech-in-noise tests discussed above, was developed by Kalikow and associates (1977) and uses a multi-talker babble as background noise over the sentence material. Cameron and Dillon (2007a:199) also employed speech as a background noise, in the form of a story being told by a talker of the same gender and accent as the one presenting the sentences.

Plomp and Mimpen (1979:44) reported using noise with a spectrum equal to the long-term average spectrum of their recorded speech material. The HINT test (Nilsson et al., 1994:1087) reports creating such a noise by filtering semi-random white noise through a filter created according to the average long-term spectrum of the recorded speech. This method was repeated by several other researchers developing or adapting tests for hearing in noise (Hällgren et al. 2006:229; Vaillancourt et al., 2005:361; Wong and Soli, 2005:279; Van Wieringen and Wouters, 2006:5). Kollmeier and Wesselkamp (1997:2413) used a slightly different technique for a German sentence test. These authors reported generating the noise by statistically superimposing all words of a monosyllabic rhyme test produced by the same speaker. Consequently, the long-term spectrum of the noise was very similar to that of the recorded speech sample.

Wilson, Carnell and Cleghorn (2007) investigated the difference between multi-talker babble (MTB) and speech-spectrum noise (SSN) as maskers for the Words-In-Noise Test (WIN). These researchers found that both normal-hearing and hearing-impaired listeners performed slightly better with MTB than with SSN. The difference was ascribed to the amplitude modulations of the MTB, which led to brief improvements in the SNR. Both types of noise clearly distinguished between normal-hearing and hearing-impaired listeners, as none of the hearing-impaired listeners had recognition performances within

the normal range (defined as the 90th percentile in normal-hearing listeners) with either noise. MTB was finally concluded to be a more appropriate masker, merely due to its face validity in representing everyday listening situations, although the findings for hearing-impaired listeners were essentially the same for the two types of noise (Wilson, Carnell and Cleghorn, 2007:528). The advantage of speech-spectrum noise is its validity as a masker for sentence materials that has been reported by multiple previous studies (Plomp and Mimpen, 1979:44; Nilsson et al., 1994:1087; Hällgren et al. 2006:229; Vaillancourt et al., 2005:361; Wong and Soli, 2005:279; Van Wieringen and Wouters, 2006:5).

Different types of noise have also been studied by Wagener and Brand (2005). These authors experimented with four different types of noise: one had the same long-term average speech or frequency spectrum (LTASS) as the sentences used for the speech stimulus; one was a random Gaussian noise with a male-weighted idealised speech spectrum that is consistent with the mean LTASS spectrum across languages; another was a three-band speech-fluctuating noise with a male-weighted idealised speech spectrum; and the fourth was a six person babble. The first two noises were stationary, and the latter two were fluctuating noises (Wagener and Brand, 2005:148).

The findings indicated that the two stationary noises (one spectrally matched to the exact material used, the other an idealised speech-weighted noise) yielded no difference in test results, whereas the fluctuating noises resulted in greater intra-subject variability. The recommendation in this regard was that a standardised interfering noise be used for different tests in different languages, provided that those languages represent the mean international LTASS. This recommendation was based on a study conducted by Byrne et al. (1994). These researchers recorded speech samples from a number of speakers for thirteen different languages and found that the LTASS across samples was so similar that it would be reasonable to use a universal LTASS for a variety of applications and languages. The authors admitted, however,

that there may be complications such as differences in some frequency-specific functions across languages and this aspect should be investigated before generalising the use of a universal type of noise to measures of speech intelligibility in any language (Byrne et al., 1994:2119).

The use of a noise specifically weighted to the speech sample used in the test still appears to be the preferred option reported in the literature (Plomp and Mimpen, 1979:45; Nilsson et al., 1994:1087; Hällgren et al. 2006:229; Vaillancourt et al., 2005:361; Wong and Soli, 2005:279; Van Wieringen and Wouters, 2006:5). This method eliminates accidental differences between the spectrum of the speaker and the noise. The different types of noise with the advantages and disadvantages of each are summarised in Table 2.1.

Table 2.1: Advantages and disadvantages of different noise types

TYPE OF NOISE	ADVANTAGES	DISADVANTAGES	REFERENCES
Multi-talker babble	Face validity in terms of representation of everyday noise	Greater intra-subject variability	Wilson, Carnell & Cleghorn (2007); Wagener & Brand (2005)
Speech-weighted noise, spectrally matched to the exact material used	Effective masker Well-documented use with sentence material	Noise needs to be generated specifically for each test	Wilson, Carnell & Cleghorn (2007); Plomp & Mimpen (1979); Nilsson et al. (1994); Hällgren et al. (2006); Vaillancourt et al. (2005); Wong and Soli (2005); Van Wieringen and Wouters (2006)
Speech-weighted noise, spectrally matched to idealised long-term speech spectrum	Universal noise can be used, with no need for creation of noise with each newly developed test	Has only been investigated in some languages, and should first be verified	Byrne et al. (1994); Wagener & Brand (2005)

2.3.3 Speaker variables

The effect of the speaker presenting the material in a speech audiometric procedure is the next variable to consider. Individual differences in vocal quality and speech production could affect the results attained during speech audiometry. Wilson, Zizz, Shanks and Causey (1990:774) have recorded

speech material from both male and female speakers and found that the sound pressure level of the material produced by the female speaker had to be increased by 11-15 dB to produce the same intelligibility scores attained with the material presented by the male speaker. These authors cautioned, however, that these findings should not preclude the clinical use of materials recorded by a female speaker, as it cannot be generalised to all male/female speakers. The findings could, however, indicate the significance of individual differences between speakers (Wilson et al., 1990:777). The effect that gender or individual differences have on the intensity level of the material can, however, be overcome by digitally adjusting these levels if material is digitally recorded (Wilson and Strouse, 1999:1337; Nilsson et al., 1994:1087). Table 2.2 indicates the gender of speakers used in the development of different speech-in-noise tests.

Table 2.2: Gender of speakers used for different speech-in-noise tests

STUDY	GENDER
Plomp & Mimpen (1979)	Female
Nilsson et al. (1994)	Male
Kollmeier & Wesselkamp (1997)	Male
Versfeld et al. (2000)	Two male, two female
Vaillancourt et al. (2005)	Male
Wong & Soli (2005)	Male
Hällgren et al. (2006)	Female
Van Wieringen & Wouters (2006)	Two male, two female
Wong et al. (2007)	Not specified
Cameron and Dillon (2007a)	Female

As shown in the table, both male and female speakers have been used for studies of this nature, and differences between genders are not specifically reported. According to Ostergard (1983:232) results attained from speakers of different genders may not compare well, especially for individuals with a high frequency hearing loss. However, in the study by Versfeld et al. (2000:1676), material presented by one of the two male speakers used in the first experiment, yielded a threshold that differed only 0.2 dB from the first female

speaker, but that differed 1.1 dB from the results obtained with the other male speaker. Likewise, the recording from the second female speaker yielded a threshold within 0.2 dB from the first male speaker, although it differed with 1.5 dB from the first female speaker (Versfeld et al., 2000:1676). The analysis of variance between the results for all four speakers revealed that the speaker had a significant effect on results (Versfeld et al., 2000:1675). These findings correlate well with the statement by Wilson et al. (1990:774) that the significance of individual differences between speakers can be generally accepted, even if gender differences cannot. Therefore, results acquired using test material presented by a specific speaker should be cautiously compared to results obtained with a different speaker, regardless of the gender of the speakers.

Additional speaker variables to consider are the speaker's dialect and pronunciation. The effect that different speakers have on speech audiometry results may be exacerbated if the speaker and listener do not have the same dialect, a problem that may be overcome by recording material in a standard dialect (Lyregaard, 1997:49). Some researchers have reported specifically selecting speakers with a standard dialect (Vaillancourt et al., 2005:360; Cameron and Dillon, 2007a:199). Additional selection criteria for a suitable speaker as stipulated in the literature are summarised in Table 2.3.

Table 2.3: Selection criteria for speaker as reported in the literature

CRITERIA	REFERENCES
Standard dialect or absence of dialectical influences	Vaillancourt et al. (2005); Cameron and Dillon (2007a)
Clear articulation	
Suitable voice quality and intonation	Versfeld et al. (2000)
Appropriate loudness and speech rate	
Pronunciation not reflective of obvious personal or social characteristics	Versfeld et al. (2000)
Pronunciation should not be breathy, untidy, dialectical, or conceited	De Villiers and Ponelis (1987)

In addition to selecting a speaker adhering to specific criteria as demonstrated in Table 2.3, some authors have provided speakers with specific instructions regarding the pronunciation of the material. This was done to ensure good quality, standard dialect recordings. Instructions included aspects such as pronouncing the material in a natural, clear manner (Versfeld et al., 2000:1672); maintaining clarity, pace and vocal effort (Nilsson et al., 1994:1087); and avoiding emphasis on key words during recordings (Vaillancourt et al., 2005:360).

In conclusion, individual differences between speakers could influence the results of speech audiometry procedures. For this reason, the use of pre-recorded material is advised. The use of digital recordings makes it possible to carefully adjust the intensity level of the speech signal, thereby eliminating unwanted loudness discrepancies (e.g. Nilsson et al., 1994:1087). In addition, the speaker used for the recordings should adhere to specific criteria and follow specific instructions.

2.4 Presentation variables for speech-in-noise tests

The presentation variables of a speech-in-noise test are related to the test procedure followed, as well as the transducer or transmission channel used during testing. This section will provide a critical overview of these aspects.

2.4.1 Presentation method

Any test aimed at determining SNR loss in a variety of subjects must have a way to prevent the ceiling and floor effects of a test score expressed in percentage (Lutman, 1997:70). This means that any test scored in percentages will always have a maximum score of 100 and minimum of 0. If a test is designed to assess individuals with a great range of capabilities in terms of speech perception, it should be able to adapt to the level of functioning of the person being tested in order to give an accurate reflection of their abilities.

One way of accomplishing this goal is through the use of a test procedure that adjusts the SNR adaptively during the test to achieve a predetermined level of performance (e.g. 50% correct). Such a method was first described by Plomp and Mimpen (1979:46). These researchers employed a fixed level of noise and adjusted the presentation level of the speech material according to the response of the subject. Following a correct response, the speech level was decreased (thereby reducing the SNR), and after a faulty response, the speech level was increased. This procedure was repeated several times until it was possible to estimate the SNR at which the subject could attain a recognition score of 50%. This is called an adaptive procedure and can be used to concentrate presentation levels in the range that yields the smallest standard deviations in threshold and slope estimates (Brand and Kollmeier, 2002:2802).

Nilsson et al. (1994:1089) also adjusted the level of the speech signal while keeping the noise level constant, as did Wong and Soli (2005:280), Vaillancourt et al. (2005:362), Hällgren et al. (2006:230), Van Wieringen and Wouters (2006:7) and Cameron and Dillon (2007a:202). Van Wieringen and Wouters (2006) used a fixed method of presentation during the development of the test where seven specific SNRs were pre-determined and a number of sentences presented to seven different subjects at one of these levels, but concluded that the adaptive procedure is preferable for clinical use of the developed instrument (Van Wieringen and Wouters, 2006:13).

It is also possible to keep the speech stimuli at a fixed level while altering the level of the noise input, as was done by Lutman and Clark (1986:1031) in a somewhat different test using word materials. The question that arises from this difference in methods is whether or not one method is more efficient or accurate than the other. Wagener and Brand (2005:155) investigated this issue and reported no difference between the adaptive method where noise is kept at a fixed level and speech level altered, and one where speech is kept

constant and the noise level altered. These authors suggested that researchers are free to choose any one of these two methods, depending on their goals and demands, since the results of these two procedures appear to be comparable.

Researchers using a fixed level of noise during the development of a speech-in-noise test have reported using intensity levels of noise ranging from 50 dB (Plomp and Mimpen, 1979:46) to 72 dB (Nilsson et al., 1994:1087). According to Wagener (2004:115) the presentation level of the noise is a non-critical factor in speech tests and can be chosen arbitrarily, as long as the noise presentation level exceeds the individual's threshold. This author reported that the threshold (SNR-50) results depended only the SNR, and not on the presentation level. This finding was confirmed by the findings of Wagener and Brand (2005:150), who found no statistically significant level effect when investigating noise level. Therefore, it seems necessary only that the noise is audible at most frequencies (Wagener and Brand, 2005:155) and does not approximate the individual's uncomfortable loudness level (Wagener, 2004:115).

2.4.2 Auditory transmission channel

The presentation of test stimuli can be conducted via a number of different methods. The two main modes of presentation are headphones and sound-field presentation (speakers). Hällgren et al. (2006:229), in development of the Swedish HINT, presented test stimuli through a loudspeaker positioned one meter in front of the subject. The advantage of sound-field testing is the application possibilities for the evaluation of amplification devices such as hearing aids or cochlear implants. However, testing in the sound-field tends to be less reliable and more variable than headphone testing, and site-specific norms must be established before using this method clinically (Vaillancourt et al., 2005:365).

There are several different methods of headphone testing suggested in the literature. The original American HINT used binaural presentation (Nilsson et al., 1994:1087), whereas a number of other researchers presented stimuli monaurally (Kollmeier and Wesselkamp, 1997:2413; Versfeld et al., 2000:1674; Van Wieringen and Wouters, 2006:5). Besides the two options of testing either monaurally or binaurally, some studies (Vaillancourt et al., 2005:363; Wong and Soli, 2005:283) described a method whereby different sound-field conditions are simulated by computer software using head-related transfer functions as measured on a KEMAR manikin². The manikin simulates the changes that occur to sound waves as they pass a human head and torso such as the diffraction and reflection around each ear. In this manner, different test conditions could be created. Vaillancourt et al. (2005:363) and Wong and Soli (2005:283) create three different conditions using this method – one with the speech and noise coming from the same direction (noise front or NF), one with speech coming from the front and noise from the right (noise right or NR), and one with speech from the front and noise from the left (noise left or NL). Both these studies reported similar results for noise right and noise left conditions, but significant differences between noise front and noise side conditions.

The noise front results of both these studies compared well with other studies in the literature (all using sentence materials and speech-weighted noise), but the noise side conditions yielded somewhat different results (see Table 2.4). The results of binaural headphone presentation for the original HINT (Nilsson et al., 1994:1087) and the sound-field presentation by Hällgren et al. (2006:229) differed from the findings of the other studies using monaural presentations (Kollmeier and Wesselkamp, 1997; Versfeld et al., 2000; Van Wieringen and Wouters, 2006). However, Plomp and Mimpen reported results from binaural presentation that correlated better with the other studies' monaural results than with the binaural findings of the original HINT. In conclusion, findings from different transmission channels do differ, but since

² The KEMAR Manikin Type 45BA can be acquired from Knowles Electronics. It is an acoustic research tool that permits reproducible measurements of hearing instrument performance on the head, and of stereophonic sound recordings as heard by human listeners.

there are a number of other variables involved (e.g. the difference in languages), these discrepancies cannot merely be ascribed to differences in the transmission channel.

Table 2.4: Results of previous studies on speech-in-noise

STUDY	TRANSMISSION CHANNEL	SNR (dB) AT 50% CORRECT
Plomp & Mimpen (1979)	Monaural (left)	- 5.6
	Monaural (right)	- 6.2
	Binaural headphones	- 7.3
	Binaural, noise uncorrelated between ears	[- 9.6]
	Binaural, diffuse noise presentation	[- 8.0]
Nilsson et al. (1994)	Binaural headphones	- 2.9
Kollmeier & Wesselkamp (1997)	Monaural headphones	- 6.2
Versfeld et al. (2000)	Monaural headphones	- 4.1 female speaker
		- 4.0 male speaker
Wong & Soli (2005)	Headphones simulating noise front, noise right and noise left conditions	- 3.9 noise front
		[-10.6] noise right
		[-10.5] noise left
Vaillancourt et al. (2005)	Headphones simulating noise front, noise right and noise left conditions	- 3.0 noise front
		[-11.4] noise side
Hällgren et al. (2006)	Sound-field (speaker in front of subject)	- 3.0
Van Wieringen & Wouters (2006)	Monaural headphones	- 7.8
Wong et al. (2007) – Mainland China MHINT	Headphones simulating noise front, noise right and noise left conditions	-4.3 noise front
		[-11.7] noise right;
		[11.7] noise left
Wong et al. (2007) – Taiwan MHINT	Headphones simulating noise front, noise right and noise left conditions	-4.0 noise front
		[-10.9] noise right
		[-11.0] noise left
Mean	(excluding noise side conditions):	- 4.79
Standard deviation	(excluding noise side conditions):	1.65
Range	(excluding noise side conditions):	4.90

Although sound-field presentation has the advantage of enabling the tester to assess listeners with hearing aids or a cochlear implant, headphone presentation is preferred to sound-field presentation in the development of a speech-in-noise test, due to the greater amount of variability present within sound-field testing (Vaillancourt et al., 2005:365). Once a test has been developed, site-specific norms for sound-field testing should be established before the test can be used as a reliable clinical procedure (Vaillancourt et al., 2005:365). As far as the choice between monaural versus binaural headphone presentation is concerned, it seems that the researcher is once again free to choose any one of these methods. However, it is important that the presentation method be specified in the final results in order to allow for accurate comparison with studies that used a similar method.

2.5 Subject variables for speech-in-noise tests

The ability of a listener to understand or recognise speech stimuli in the presence of a background noise is influenced by a complex combination of factors, both in terms of the test procedure, and internal to the listener. The internal factors affecting speech recognition in noise can be divided into two groups. The first group of factors are related to peripheral hearing sensitivity, or more specifically, hearing impairment. The second group consists of factors that are not directly related to hearing sensitivity. During the development of a test that intends to measure this ability, researchers should attempt to control these factors as far as possible. In the clinical application of a developed test, these factors can assist in the interpretation of test results. This section will focus on the internal subject characteristics that affect speech recognition, especially in the presence of noise.

2.5.1 Peripheral hearing

An impairment in peripheral hearing sensitivity will lead to a loss of speech intensity and therefore make soft speech sounds less audible, as is the case

in a conductive hearing loss (Gelfand, 2001:176). However, a sensorineural hearing loss not only attenuates the speech signal, but also causes distortion thereof, thereby causing significant problems for the individual with such a hearing loss (Gelfand, 2001:175). In the development of a test that assesses speech recognition, it is therefore necessary to ensure that subjects used during the developmental process have normal hearing.

The dilemma in selecting normal-hearing subjects lies in the definition of the term “normal hearing”. In the strictest sense of the term, normal hearing implies a completely unimpaired development of hearing function and no history of any pathological insult to the ears or auditory system (Blandy and Lutman, 2005:435). This would disqualify a great number of people, as virtually all individuals might have been at risk of some form of auditory damage during their lives. To overcome this problem, the International Organization for Standardization (ISO) has introduced the concept of an otologically normal person – defined as “a person in a normal state of health who is free from all signs and symptoms of ear disease and from obstructing wax in the ear canal, and who has no history of undue exposure to noise” (International Organization for Standardization, 1991). To prevent the effect that age might have on this system, the hearing of young adults (18 to 25 or 30 years old) are generally used as a reference or baseline (Blandy and Lutman, 2005:435).

Previous studies aimed at developing a speech-in-noise test using sentence materials have differed slightly in their selection criteria as far as peripheral hearing sensitivity is concerned. Table 2.5 provides a description of the criteria specified by these studies. From this comparison it is clear that the most common selection criteria specify that subjects are required to have pure tone thresholds of 15 dB HL or better for octave frequencies from 250 to 8000 Hz. The additional criteria stipulated by Vaillancourt et al. (2005:360), namely a normal otoscopic examination, normal tympanograms and negative otologic history are more conservative, but also show greater concurrence with the

ISO's definition of an otologically normal person (International Organization for Standardization, 1991). Therefore, it is recommended that these criteria be included in the selection of subjects for the development of a speech-in-noise test.

Table 2.5: Auditory selection criteria of previous studies

AUTHORS	AUDIOLOGICAL CHARACTERISTICS OF SUBJECTS
Plomp & Mimpen (1979)	"Normal hearing" (only specification)
Nilsson et al. (1994)	Thresholds \leq 15 dB HL from 250 to 8000 Hz
Versfeld et al. (2000)	Thresholds \leq 15 dB HL from 250 to 8000 Hz
Wong & Soli (2005)	"Normal hearing"
Vaillancourt et al. (2005)	Thresholds \leq 15 dB HL from 250 to 8000 Hz; normal otoscopic examination; normal tympanograms; negative otologic history
Hällgren et al. (2006) 1 st phase (Equating sentence difficulty)	Self-reported normal hearing
Hällgren et al. (2006) 2 nd phase (Inter-list reliability)	Thresholds \leq 15 dB HL from 250 to 8000 Hz
Van Wieringen & Wouters (2006)	Thresholds $<$ 20 dB HL from 125 to 8000 Hz
Wong et al. (2007)	Thresholds \leq 15 dB HL from 250 to 8000 Hz
Cameron and Dillon (2007a)	Thresholds \leq 15 dB HL from 500 to 4000 Hz; thresholds \leq 20 dB HL at 250 and 8000 Hz; normal Type A tympanogram and ipsilateral acoustic reflex (1000 Hz) at 95 dB HL on day of test

2.5.2 Age

Age is a subject characteristic that is indirectly related to peripheral hearing. According to the National Institute on Deafness and Other Communication Disorders in the United States (2007), the incidence of hearing loss increases with age, with approximately 30% of adults over 65 having a hearing loss, and between 40 and 50% over the age of 75 suffering from a hearing impairment. Barrenäs and Wikström (2000:569) investigated the effect of hearing loss and age on speech recognition scores in both quiet and noise. Their findings indicated that age had no influence on recognition scores if hearing was normal, but did influence the results in the presence of a hearing loss. By implication, the age of normal-hearing subjects used in the development of a

speech-in-noise test should not influence the outcomes, but in the clinical administration of the procedure, a patient's age could interact with his hearing loss to influence the results.

However, there may be another reason why it is important to control the age of subjects in developing test material. Among the collection of factors influencing speech recognition that are not directly related to peripheral hearing sensitivity are cognitive aspects such as deficits in memory, problem solving or reasoning (Crandell, 1991:102S). Theories on the influence that these factors have on speech recognition have originated mainly to explain the difficulties that elderly listeners have in this regard. Therefore, in order to control this variable during the development of a speech-in-noise test without having to formally evaluate the cognitive abilities of all subjects, researchers could specify selection criteria in terms of the age of subjects. In this way, the effect of aging on the cognitive abilities of subjects may be controlled.

Furthermore, it could be specified in the selection criteria that subjects should not suffer from learning disabilities or cognitive impairment. A useful fail safe may be to specify that subjects must have completed Grade 12 in a mainstream school. Such a criterion would be similar to that of Vaillancourt et al. (2005:360), who specified that their subjects were required to have finished post-secondary education in French (the language in which the test was developed). Other researchers have not set specific criteria in terms of the cognitive abilities of their subjects, but have nevertheless found small standard deviations between subjects in their final results (Nilsson et al., 1994:1090; Hällgren et al., 2006:231; Versfeld et al., 2000:1679). These researchers did, however, specify the age range of their subjects, which could have at least controlled for the effect of aging on cognition. Table 2.6 indicates the age ranges specified by previous researchers. It is evident from these specifications that the age of 45 years is the highest upper limit of age reported, but the mean age of subjects should be much lower than that.

Table 2.6: Age ranges of subjects in previous studies³

AUTHORS		RANGE	MEAN AGE
Plomp and Mimpen (1979)		Not specified	
Nilsson et al. (1994)	Phase I	17-45 years	24 years
	Phase II	18-43 years	26.8 years
Kollmeier and Wesselkamp (1997)	Phase I	19-31 years	
	Phase II	22-36 years	
	Phase III	22-36 years	
Versfeld et al. (2000)	Phase I	20-53 years	26 years
	Phase II	18-43 years	22 years
	Phase III	18-26 years	22 years
Vaillancourt et al. (2005)		18-45 years	
Wong and Soli (2005)		Not specified	
Hällgren et al. (2006)	Phase I	19-40 years	24 years
	Phase II	18-30 years	21.3 years
Van Wieringen and Wouters (2006)		20-25 years	
Wong et al. (2007)	Mainland China subjects	Not reported	22 years
	Taiwan subjects	Not reported	26.9 years

2.5.3 Central Nervous System

Populations with language abilities that are insufficient to permit the use of conventional adult speech audiometry include infants and young children; hearing-impaired persons with reduced verbal language skills; individuals of whom the first language differs from the language of the test; mentally retarded persons; and people with aphasia (McLaughlin 1980:253). For the development of a speech-in-noise test these populations should be excluded from the subject pool, since the utilisation of linguistic information stored in the individual's memory make up an important part of understanding sentences (Kalikow et al., 1977:1337).

A number of previous studies aimed at developing a speech-in-noise test with sentences have therefore reported that their subjects were required to be native speakers of the test language (Nilsson et al., 1994:1087; Vaillancourt et

³ Results of Cameron and Dillon (2007a) excluded, as their test was exclusively developed for children

al., 2005:360; Hällgren et al., 2006:229; Wong et al. 1007:70S). To clearly define a “native speaker”, selection criteria may be formulated specifying that the test language should be the language primarily used on a daily basis or the language spoken most often at home (Vaillancourt et al., 2005:360; Statistics South Africa, 2001:5). Furthermore, the requirement of having completed post-secondary education (Grade 12) in the language of the test could be included as a selection criterion (Vaillancourt et al., 2005:360).

An additional advantage of including only subjects that have completed post-secondary education in a mainstream school is that it would exclude from the sample subjects with mental retardation, thereby controlling the effect of cognitive abilities (McLauchlin, 1980:253; Vaillancourt et al., 2005:360). Besides cognitive abilities, central auditory lesions tend to affect an individual’s ability to understand speech, especially in difficult listening conditions (Crandell, 1991:102S). Difficulties in auditory processing have been found in children with language-learning problems and people with known lesions to the central auditory system, but also in individuals whose only complaint was an apparent inability to hear well in difficult listening situations (Neijenhuis, Stollman, Snik and Van den Broek, 2001:69). The significant self-reported listening difficulties of individuals with normal peripheral hearing are often the first indication that these patients may have an Auditory Processing Disorders (APD) (Neijenhuis et al., 2001:69; Bellis, 2003b:10). This indicates that it may be possible to screen for this disorder by means of a questionnaire that should include questions on hearing in noise (Meister, Von Wedel and Walger, 2004:436).

Auditory processing difficulties are also known to be associated with a history of persistent otitis media with effusion (Bellis, 2003a:134). Children with a history of otitis media have been found to have abnormally reduced masking level differences, which affects their ability to extract signals from noise (Geffner, 2007:39), and the conductive hearing loss caused by the infection could have a long term effect on their auditory processing abilities (Bamiou,

2007:267). In addition, neurologic disease, neurosurgery, traumatic brain injury and aging could cause APD in adults (Bellis, 2003b:10). Therefore, it is important to control for these variables in the development of a new test for speech recognition in noise by questioning potential subjects on these issues and excluding individuals with these risk factors from the study.

Another variable, which is partly related to the subject and partly to the test procedure, is the extent to which a subject can guess the correct response. Burke and Nerbonne (1978:89) found that guessing has a significant effect on test results, affecting the traditional SRT with an average of 4.2 dB. This emphasises the need of clear and consistent instructions, also in terms of the amount of guessing. Burke and Nerbonne (1978:90) suggests that consideration be given to instructing patients not to guess at all, but to repeat only stimulus items of which they are certain. However, these authors do admit that this method will still not exert perfect control over the amount of guessing exercised by each subject.

Many previous researchers that developed a test for speech recognition in noise have encouraged their participants to guess and have still managed to obtain reliable results from their subjects (Plomp and Mimpen, 1979:44; Nilsson et al., 1994:1087; Kollmeier and Wesselkamp, 1997:2413; Versfeld et al., 2000:1674; Vaillancourt et al., 2005:361; Wong and Soli, 2005:281; Hällgren et al., 2006:230; and Van Wieringen and Wouters, 2006:6). In addition, allowing subjects to guess could improve the extent to which the results reflect a listener's ability to cope with background noise in their daily lives, where listeners are free to revert to guessing in order to understand conversations.

2.6 Response variables for speech-in-noise tests

2.6.1 Response channel

In a test where the subject is required to identify or recognise a particular speech item, they may be asked to repeat aloud what they heard, or asked to write down their response (Lutman, 1997:70). If the subject is required to repeat aloud what was heard, it should be ensured that the tester is able to hear clearly what is being said. This is especially important in closed-set tests where the different options closely resemble each other. Whatever the nature of the test material, ensuring an optimal acoustic environment such as a sound-treated booth could enhance transmission of the response.

Numerous previous studies aimed at developing speech-in-noise tests all instructed their subjects to repeat aloud what was heard and encouraged them to guess (Plomp and Mimpen, 1979:44; Nilsson et al., 1994:1087; Kollmeier and Wesselkamp, 1997:2413; Versfeld et al., 2000:1674; Vaillancourt et al., 2005:361; Wong and Soli, 2005:281; Hällgren et al., 2006:230; and Van Wieringen and Wouters, 2006:6). The disadvantage of having the subject repeat the sentence verbally, is that it may be possible for the tester to misinterpret or mishear the response, especially if a correct response is anticipated. Furthermore, having the subject respond verbally, leaves the test administrator without a written record of the response that could have been used for further analysis or review at a later stage.

Having a written copy of subject responses could be especially valuable in the development of a new test, as different scoring methods could be experimented with after testing, and error patterns could be analysed. However, written responses could also be misinterpreted by the tester, and spelling mistakes could cause additional distortion of the response. A possible solution or middle ground should thus be for the subject to repeat stimuli aloud (in order to prevent distortion through spelling mistakes or unclear handwriting), but for some written record to be kept by the test administrator to make later analysis of responses possible. This could be done if the tester

had a form containing a text version of the stimuli and recorded the subject's responses on the form.

2.6.2 Scoring method

The method of scoring for a speech-in-noise test is partly determined by the manner in which subjects' responses are compared to the stimulus material. In this regard, two different issues can be identified. The first is the method of scoring followed during the development of the test, and the second is the scoring method applied during the clinical application of the test, once it is in the final format.

With the development of the American HINT, scoring during the initial phases of test development was done on a word-by-word basis and only exact repetitions were accepted as correct. The word-by-word scoring enabled the researchers to assign a percentage value to the correctness of each sentence's repetition by calculating the percentage of words repeated correctly under a specific listening condition. In this way, it was possible to compare the difficulty of sentences by comparing the percentage score each sentence yielded at a fixed SNR. This method could provide a basis on which sentences can be eliminated or adjusted in order to yield a final collection of equally intelligible sentences (Nilsson et al., 1994:1087; Vaillancourt et al., 2005:361).

During the development of the HINT, the word-by-word scoring method was replaced in the final format of the test by scoring the whole sentence as correctly or incorrectly repeated (Nilsson et al., 1994:1088) and scoring criteria were relaxed to allow for minimal variations in articles and verb tenses, e.g. a/the or are/were substitutions (Nilsson et al., 1994:1088). The "whole sentence" scoring method can be used for the adaptive measurements of SNR-50 thresholds in the final phase of test development (Vaillancourt et al., 2005:361). During this phase, a list of sentences is presented to the listener, who repeats them to the test administrator. The administrator has to make a

quick decision on the correctness of the sentence (hence the simple right/wrong scoring method), and according to this determines the presentation level of the next sentence. In other words, if the listener repeats a sentence correctly, the SNR is decreased. If the sentence is repeated incorrectly, the SNR is increased or improved. This is called an adaptive up-down presentation strategy (Nilsson et al., 1994:1089).

Vaillancourt et al. (2005:361) used the same method of word-by-word scoring for test development and whole sentence scoring in the final format, with the same rationale. These authors also relaxed their criteria in the later phases of the project. Acceptable substitutions in sentences were determined by reviewing the original HINT material, and discussing the substitutions most frequently made by subjects during initial trials. Pronoun gender variations as well as article variations were considered to be acceptable substitutions. As with the original HINT, the acceptable substitutions were indicated in parentheses on the scoring form, e.g. “(He/she) is writing (a/the) book” (Vaillancourt et al., 2005:362). Hällgren et al. (2006:230) followed this same method, and eventually allowed for variations in verb tense, articles and singular versus plural nouns.

The limitation of word-by-word scoring during test development is that it constitutes a rough indication of the performance of subjects on each sentence, especially when using short sentences. A sentence consisting of only four words, for example, can only receive 25, 50, 75 or 100%. Furthermore, it does not give any credit for multi-syllabic words in which a subject made even the slightest mistake, e.g. confusing singular with plural. An alternative to address this limitation could be the use of syllable-by-syllable scoring. In this way, a more detailed impression of performance on each sentence could be acquired. In addition, subjects would receive some credit for a multi-syllabic word where only a small mistake unrelated to the main content of the sentence (such as a plural/singular substitution) was made.

This method may be especially valuable in the development of sentence tests in languages where there is a tendency in the spelling rules to write conjunctions as one word (the so-called conjunctive method) as opposed to the English tendency to write conjunctions as two words (the disjunctive method) (Carstens, 2003:232). In these languages, there may be many multi-syllabic conjunctions that could receive a more precise scoring if syllable scoring is used. Although this method has not been documented by any previous researchers, it may be valuable to investigate as an alternative method.

2.7 Test performance variables for speech-in-noise tests

Speech-in-noise tests, just like other psychometric tests, should adhere to certain standards in order to ensure their accurate performance as measures of speech perception (Lucks Mendel and Danhauer, 1997:7). In order to serve their purpose as an objective and accurate measure of a listener's speech perception abilities, these tests must be sensitive, specific, reliable and valid (Ostergard, 1983:222). As indicated in Figure 2.2, these criteria or variables are interrelated. For instance, for a test to be valid, it must be reliable, but a measure can be reliable without being valid (Lucks Mendel and Danhauer, 1997:11). In order for a test to be sensitive, it must be both valid and reliable (Lucks Mendel and Danhauer, 1997:7). Sensitivity and specificity are interrelated in a reciprocal fashion – gains in one will inadvertently lead to losses in the other (Ostergard, 1983:225).

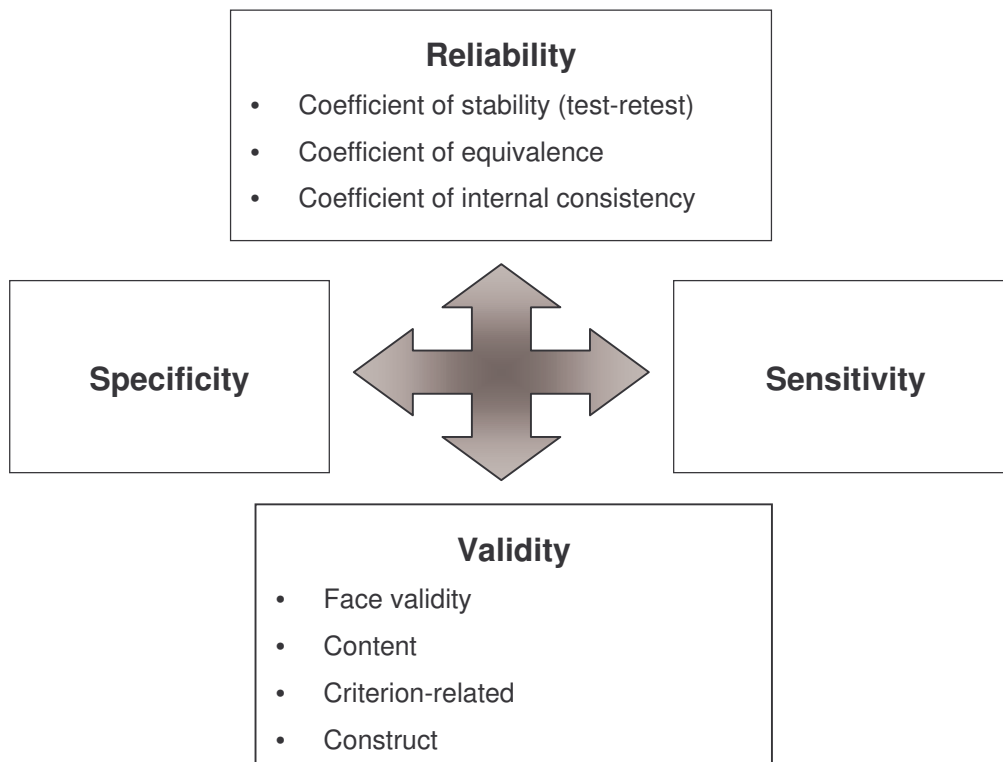


Figure 2.2: Factors influencing test performance

Due to the complex interrelation between these variables, it is important to attend to all these factors in the development of a new speech perception test. The standards that a test using sentence materials for determination of a speech recognition threshold should meet are exceptionally high (Plomp and Mimpen, 1979:43). The reason for this is that these tests are often aimed at detecting very subtle changes in the threshold that could be induced by a small degree of hearing loss, or a small adjustment made to a hearing aid's settings. The performance of such a test is basically influenced by the four factors shown in Figure 2.2. These aspects will be critically discussed in the following section.

2.7.1 Validity

The validity of a test is determined by the extent to which the test can achieve its aims or measure what it is supposed to measure (Ostergard, 1983:223; Lucks Mendel and Danhauer, 1997:8). Different types of validity are

mentioned in existing literature. These include face validity (Lucks Mendel and Danhauer, 1997:8), content validity, criterion-related validity and construct validity (Ostergard, 1983:223).

2.7.1.1 Face validity

Validity relates to the extent to which a test achieves its aims (Ostergard, 1983:223) and is therefore largely determined by the goal of the test. Mackersie (2002:392) identifies two possible basic goals for speech tests. The first is to assess a person's auditory capacity without involving cognitive and linguistic factors. The second possible goal is to describe a listener's ability to recognise spoken language using materials that do involve higher-level cognitive and linguistic skills. The latter type of test would typically give a better indication of a person's everyday functioning, since everyday auditory stimuli usually come in a form that does require cognitive and linguistic involvement. It could therefore be said that such a test appears to have a high degree of validity for estimating an individual's ability to deal with typical speech stimuli. This apparent validity is termed "face validity" – defined as "the extent to which a test instrument appears to measure what it is supposed to measure" (Lucks Mendel and Danhauer, 1997:8). Although face validity has a strong influence on the acceptance of a test, it can be misleading if there is not stronger evidence of its real validity, and other types of validity should also be considered.

2.7.1.2 Content validity

Content validity is related to the extent to which the test items reflect the behaviour of interest (Lucks Mendel and Danhauer, 1997:8). If inferences are therefore to be drawn about a person's ability to cope with typical everyday speech stimuli, a test with high content validity will consist of test items that closely resemble these typical stimuli. This concept also aligns with the term "ecological validity" explained in the first chapter, which indicates the extent to which the results of a test reflect speech perception abilities in natural listening environments (Mackersie, 2002:395). In this regard, the ideal stimuli

for a test that aims to resemble everyday speech would probably be conversational speech. However, the complexity of such stimuli would make scoring and administration extremely difficult. In terms of existing, workable tests, those using real sentences as stimuli have a higher degree of content validity (if the aim is inferences about everyday functioning) than those using single words or syllables (Nilsson et al., 1994:1086).

The term “cultural validity” has also been used to describe the extent to which test material is appropriate for a particular cultural group (Roets, 2005:67; Pakendorf, 1998:2). This could be related to content validity, in the sense that test material must reflect vocabulary and test items familiar to, and typically used by, the cultural group targeted by the test. To achieve this type of validity, previous researchers developing a sentence recognition test have submitted their sentence material to native speakers of the test language for a rating of the naturalness of the sentences (Nilsson et al., 1994:1086; Vaillancourt et al., 2005:360; Hällgren et al., 2006:228; Wong and Soli, 2000:278).

The presence or absence of noise as part of the stimulus also exerts an influence on content validity. If a test’s content is intended to reflect typical everyday situations, stimuli must be presented in the presence of some degree of background noise. However, the types of noise that individuals are exposed to in their daily routines vary considerably and it would therefore not be possible to compile a test with the exact type of noise every person faces on a day to day basis. Also, a highly variable noise would cause some test items to be more difficult than others and influence the reliability of the test. In order to achieve some degree of “content validity” in terms of noise without compromising the reliability of the test, stimuli should be presented in the presence of noise, but this noise should be of a controlled, known intensity and frequency (Wagener and Brand, 2005:155). According to Wilson, Carnell and Cleghorn (2007:528), multi-talker babble noise is more representative of the type of everyday noise that listeners find problematic than speech

spectrum noise, but this noise type has been found to increase intra-subject variability in the results of a speech-in-noise test (Wagener and Brand, 2005:155).

2.7.1.3 Criterion-related validity

The third type of validity is criterion-related validity (Ostergard, 1983:223). This can also be termed predictive validity, and relates to the correlation between the test's score and other measures of the same behaviour (Lucks Mendel and Danhauer, 1997:8). Previous developers of sentence recognition tests do not commonly report on the criterion validity of the developed measures, although Wilson, McArdle and Smith (2007) have conducted a study to compare performance of English listeners with normal hearing and hearing-impaired listeners on four commonly available speech-in-noise protocols (HINT; Bamford-Kowal-Bench Speech-In-Noise test or BKB-SIN; Quick Speech-In-Noise test or QuickSIN; and the Words-In-Noise test or WIN). However, if a test is being developed in a language where no established tests of the same behaviour exist, it is not possible to evaluate the criterion validity.

2.7.1.4 Construct validity

In cases where criterion validation is not possible, the evaluation of construct validity is recommended in order to verify that the data collected by the test correlate with the theoretical constructs underlying it (Ostergard, 1983:223). In the case of a test measuring sentence recognition in noise, this applies to the extent to which the results relate to the theory underlying speech perception in noise. An example of such a theory is the principle that the ultimate intelligibility of a speech signal depends not only on whether it is audible for the listener, but also on the degree to which the auditory system can make use of the signal (Gatehouse and Robinson, 1997:79). Due to this effect, patients with similar audiograms may have vastly different abilities to understand speech in noise (Killion and Niquette, 2000:50). A test of speech recognition in noise could provide a means to quantify this ability.

In order to verify the construct validity of a test, it would have to be applied to a population showing a deficit in this area, as done by Van Wieringen and Wouters (2006:12). These researchers applied their developed sentence and numbers tests to a group of cochlear implantees, and found the material a valid and feasible method of assessing speech recognition in this population. However, applying the test to a hearing-impaired population introduces a great number of new variables to the development of a speech recognition test, as a group of hearing-impaired subjects may be of different ages than the normal-hearing subjects used, and might have a great variety of audiograms and auditory difficulties beyond peripheral hearing loss. For this reason, it may be simpler to rather simulate a hearing loss in the same normal-hearing group of subjects already partaking in the study. Past researchers have followed this method to test hypotheses by simulating certain characteristics of a hearing loss (Stuart, Phillips and Green, 1995:659; Scott, Green and Stuart, 2001:439). In the development of a speech-in-noise test, this method could enable researchers to compare findings of each subject with and without the simulated loss, thereby reducing the number of variables affecting findings. This possibility will be discussed in further detail under section 2.7.2.3.

2.7.2 Reliability

In order for a measure to be valid, it must be reliable (Lucks Mendel and Danhauer, 1997:11). Reliability can be defined as the consistency of a test's results across a series of different observations (Ostergard, 1983:224). This means that the test results should stay consistent if the test is repeated; either by the same test administrator, or by a different administrator. Three types of reliability are identified: a coefficient of stability; a coefficient of internal consistency; and a coefficient of equivalence (Ostergard, 1983:224-225). Since all three of these types should be considered in the development of a speech-in-noise test to ensure reliability, a discussion of each will follow.

2.7.2.1 Coefficient of stability

The coefficient of stability refers mainly to the test-retest correlation of a given measure (Ostergard, 1983:224). This pertains of course to enduring features, and not those that are meant to change over time. In the case of a speech-in-noise test, it would be expected that a person's performance on the test remains stable, provided that their peripheral and central hearing remained constant. One way to evaluate this would be to test and retest otologically normal individuals (with normal hearing), since their hearing threshold and speech recognition abilities could be expected to remain stable across time.

The dilemma of evaluating this aspect in speech audiometry is that repeated exposure to the material could largely improve performance, since speech materials become less difficult as they are reused (Nilsson et al., 1994:1085). Listeners are therefore expected to perform better during a retest due to the increased familiarity of the material (learning effect), but it would be impossible to say how much of the improvement was due to this learning effect, and how much could be ascribed to poor test-retest reliability of the measure itself.

This learning effect could be limited by familiarising subjects with the test material beforehand, as has been suggested for determining the spondee threshold (Ostergard, 1983:231). The recommended protocol in this case is a single reading of the word list at a supra-threshold level before the commencement of the test (Rupp, 1980:91). However, this does not appear to be common practice in the development of tests using sentence material. This could be due to the volume of the test material, which would make the familiarisation a lengthy process and would make it fairly difficult for a listener to remember the material after hearing it only once.

The developers of the original HINT (Nilsson et al., 2004) and many of their followers who adapted the test in other languages did not report on the evaluation of test-retest reliability of their developed measures. Hällgren et al.

(2006:231), however, did assess the test-retest reliability of their speech-in-noise test by evaluating the same subjects with the same lists in the same order after one week. These authors did not familiarise subjects with the material before testing, and found only a small improvement of less than 1 dB on the mean SNR during the retest. Their findings suggest that test-retest reliability can be measured reliably without familiarising subjects to the material beforehand, and that a small improvement in mean SNR can be expected with the second test.

Developers of other speech-in-noise tests that used slightly different material, noise types or target populations have also reported on the evaluation of test-retest reliability. The “Listening In Spatialized Noise-Sentences Test” or LISN-S was developed by Cameron and Dillon (2007a) for use in a paediatric population and uses continuous discourse (distracter stories) as background noise. The test-retest reliability of this measure was evaluated by re-testing 46 of the children that participated in the normative study after two months (Cameron and Dillon, 2007b). These researchers were able to successfully assess the reliability of the LISN-S by analysing the differences and correlations between the results of the first and second evaluations (Cameron and Dillon 2007b:152).

2.7.2.2 Coefficient of internal consistency

A second type of reliability is called the coefficient of internal consistency, also called the split half test (Ostergard, 1983:225). This refers to the extent to which results yielded by half of the items in a test compare to results attained by the other half of the items. For example, if 25 words were randomly selected from a list of 50 words and applied to a subject, it should yield similar results to the application of the other 25 words to the same subject. However, the split-half test cannot be applied in this way to a sentence test where sentences are grouped into lists of ten and the entire list is used to determine one value (the SNR-50), as described by Nilsson et al. (1994:1090). Nevertheless, it may be possible to compare the thresholds obtained by a

certain combination of lists with the thresholds from a different combination of lists, although this has not been previously reported. Alternatively, internal consistency could be assessed by conducting an analysis of variance on the effect of list on the threshold value, as reported by Vaillancourt et al. (2005:363).

2.7.2.3 Coefficient of equivalence

The third type of reliability is the coefficient of equivalence, (Ostergard, 1983:224). There are two dimensions to this form of reliability. Firstly, the test should yield similar results when administered to the same subject by two different testers. In the case of speech audiometry, the variability of monitored live voice procedures often diminishes the reliability of testing, due to the variability of the stimulus (Konkle and Rintelmann, 1983:7). Therefore, pre-recorded stimuli are recommended as a standard procedure to reduce this variability (Ostergard, 1983:224). Previous studies reporting on the development of speech-in-noise tests such as the HINT (Nilsson et al., 1994) therefore all report using pre-recorded sentences that were scaled to have the same average intensity. These studies, however, do not report on the comparison of results between different test administrators.

The second dimension of the coefficient of equivalence is the stability of results across different forms or lists of the same test, also called inter-list reliability (Ostergard, 1983:225; Nilsson et al., 1994:1089). The equivalence in difficulty between lists is of the utmost importance in the development of a speech-in-noise test. The reason for this is that the test may have to be repeatedly applied to the same individual, as tests of this kind are often used for monitoring progress in rehabilitation or evaluating amplification efficiency (Rupp and Stockdell, 1980:5). However, due to the redundancy of sentence materials, stimuli are too easily recognised if repeated (Owens, 1983:359). Therefore, a sentence test must consist of a large enough collection of items or lists that the same person can be tested repeatedly without the familiarity of stimuli affecting test-retest reliability. This means that applying two different

lists to the same person should yield similar results so that list difficulty remains a controlled variable. In this way, the tester can be sure that what is really being measured is a difference in speech recognition abilities (due to adjustments made to the hearing aid, for example), and not a difference between two lists. Inter-list equivalence is usually determined by comparing the mean score for each list across subjects with the overall mean, i.e. the average threshold for all lists across all subjects (Nilsson et al., 1994:1090; Wong and Soli, 2005:285; Vaillancourt et al., 2006:363; Hällgren et al., 2006:231; Wong et al., 2007:72S) or looking at the standard deviation of the mean scores across subjects (Kollmeier and Wesselkamp, 1997:2416; Versfeld et al., 2000:1680).

In order to achieve this equivalence, past researchers have applied a number of different methods. Plomp and Mimpen (1979:45) arranged sentence material into lists based on the distribution of different phonemes in each list, using a computerised process. Phoneme occurrences or frequencies were determined for each list, and sentences were interchanged between lists in order to balance phoneme occurrence across lists. This process proved successful in balancing the phoneme content of the different lists, but required five hundred interchanges before a significant difference was made, making this a very long and tedious method.

Nilsson et al. (1994:1088) also based their list composition on phonemic distribution. After phonetic transcription of the material, a similar trial-and-error process was used to match the phonemic distribution of each list with the distribution of the entire collection. Equivalence between lists was then evaluated by applying it to test subjects and comparing the scores yielded by different lists. Hällgren et al. (2006:229) followed this same method for the Swedish HINT, as did Vaillancourt et al. (2005:362) for the French-Canadian version of the test and Wong and Soli (2005:282) for the Cantonese HINT. The only notable difference between these studies was that the American

HINT initially arranged sentences into lists of twelve, whereas the other studies had only ten sentences per list.

Kollmeier and Wesselkamp (1997:2414) followed a more complex method to attain inter-list equivalence. These researchers assembled lists of 10 sentences each according to the mean discrimination function of all the sentences in the list, the number of words and phonemes in each list as well as the frequency distribution of phonemes among lists. This was done by means of a numerical optimisation procedure. Although this procedure seems a prudent way of assuring list equivalence, it requires substantial effort and mathematical expertise in its execution. The high degree of list equivalence attained by researchers who merely balanced lists in terms of phonemic distribution, makes the necessity of such a complex method questionable.

The value of phonetically balancing material has also been questioned in the literature. In 1964, Tobias suggested that clinical and experimental experience indicates phonetic balance to be an “interesting but unnecessary” aspect of speech audiometric tests (Tobias, 1964:99). Carhart (1970:229) agreed that precise phonetic balance appears to be of little clinical importance. Aspinall (1973 in Bess, 1983:166) experimented with four different word lists of varying degrees of phonetic balance (one list consisted of randomly selected words) and found no significant difference in scores when applied to a group of subjects with sensorineural hearing loss. This correlates with the more recent findings of Martin, Champlin and Perez (2000:489) who used four phonetically balanced lists and four lists consisting of randomly selected words to evaluate both normal-hearing subjects and subjects with mild-to-moderate sensorineural hearing losses. These authors found no clinically significant differences in results, thereby adding strength to the argument against phonetically balanced lists.

The practice of balancing the phonetic content of sentence lists is not motivated or justified in any of the studies mentioned earlier, and therefore

seems to be merely a manner of ensuring list equivalence. If this is the case, there may be a simpler way of compiling equivalent sentence lists. Since lists are usually compiled after sentence materials have already been subjected to perceptual evaluation at different SNRs (Nilsson et al., 1994), there should already be data available to predict the SNR-50 that each sentence will yield, along with the expected intelligibility slope. If sentences were to be grouped together according to previous findings for performance at two or three different SNRs, it should be possible to calculate the expected slope and SNR-50 for a group of ten sentences according to the mean performance of all the sentences in the list. Sentences could then be exchanged between lists until all the lists yield an equivalent slope and SNR-50. Such a method would be much less time-consuming than the phonetic transcription and balancing of a long list of sentences.

However, there is an additional factor related to phonetic content that could affect the reliability of the test, although it is not mentioned by any of the researchers using phonetic content to attain list equivalence. The effect that a hearing loss has on an individual's speech recognition depends partly on the frequency range affected by the hearing loss. This is because the high frequency components that characterise some of the consonants are lost (Davis, 1970:92). It could thus be said that the hearing loss acts as a type of "frequency filter". If two different lists in a speech-in-noise test therefore contain different amounts of a phoneme that is particularly difficult for a person with a hearing loss due to its specific frequency, the results will differ and the lists are therefore not equivalent. In other words, lists should be balanced in terms of phonetic content in order to ensure that they are all equally difficult for listeners with a hearing impairment.

If lists were to be compiled according to simpler criteria such as the predicted intelligibility slope, and not according to phonetic content, they would have to be validated on a population with a hearing loss at specific frequencies, to ensure that list equivalence is not affected by the lack of phonetic balance.

However, the introduction of a group of individuals with hearing impairment in the study would imply the introduction of a great number of extra variables. All of the subject variables discussed previously will be affected, since age, peripheral hearing and auditory processing abilities may all be quite different in the hearing-impaired group when compared to the group of normal-hearing listeners used in the development of such a test.

In order to reduce the number of new variables introduced, it is possible to simulate a high frequency hearing loss in subjects in order to evaluate the equivalence of lists that are not phonetically balanced. In this way, the same subjects could be used and a comparison made of inter-list variability with and without the simulated hearing loss or “frequency filter” effect in the same individual. Stuart et al. (1995) used this method of simulation when studying the hypothesized effect of a high frequency hearing loss on temporal resolution. These authors used a low-pass filter with a roll-off slope of 48 dB per octave to simulate a high frequency hearing loss (Stuart et al., 1995:660).

Although such a method by no measure simulates all the auditory effects of a cochlear hearing loss, it should be sufficient to examine the frequency filter effect that the hearing loss could have on list equivalence. This method will allow researchers to examine the intra- and inter-subject variability both across the different lists and across the filtered and unfiltered conditions. If the variability between lists increases significantly when the material is filtered, the developed test may not have sufficient inter-list reliability for application in a hearing-impaired population.

Due to the effect of reduced bandwidth, filtering the speech signal could, however, influence the inter-list variability even if the lists are phonetically balanced. The developers of the American HINT (Nilsson et al., 1994), postulated that a reduction in bandwidth of the signal and/or the noise will influence the threshold results for a sentences-in-noise test, in that the SNR would need to be increased as the bandwidth is reduced. They also

anticipated, however, that at some point the threshold will be influenced more by the bandwidth than by the level of the speech, since response biases and guessing will then have an increasing effect on the reliability of the measurements (Nilsson et al., 1994:1093). They undertook an investigation into the bandwidth required to obtain reliable thresholds, as they hypothesized that hearing impairment could reduce audible bandwidth and therefore affect the measurement reliability, despite the fact that the material was phonetically balanced.

Their findings indicated that elimination of 4- and 8-kHz octave bands merely elevated thresholds, but when bandwidth dropped below about 2 kHz, the reliability of the thresholds was substantially degraded (Nilsson et al., 1994:1095). Therefore, it seems that filtering of the speech signal will affect inter-list reliability even when the lists are phonetically balanced, and such lists should also be subjected to “filtered” testing. Should a different method of list compilation be followed (such as compiling lists according to intelligibility slopes), it would be of great value to investigate the coefficient of equivalence of a filtered version of these lists and compare these findings to the results of a filtered version of phonetically balanced lists.

2.7.3 Sensitivity and specificity

The sensitivity of a test refers to the rate of correct identification of affected individuals, that is, how accurately it identifies all individuals who have a given disorder (Roush, 2001:20). In other words, a test that is 100% sensitive will not miss any individuals with a deficit in a given area. Sensitivity and specificity influence each other in a reciprocal manner – a test that is 100% sensitive will most probably have a low specificity and therefore a poor efficiency (Ostergard, 1983:226). Specificity refers to the rate of correct classification for unaffected individuals, i.e. accurately identifying persons who do not have the condition screened for (Roush, 2001:20). In other words, a test that is 100% specific will not falsely identify any unaffected (healthy) individuals as having the disorder that was tested for. This would make for a

very efficient test, as no unnecessary referrals would be made. Unfortunately, a test that is so specific, will inevitably miss individuals who do have the disorder, that is, its sensitivity will be affected. When working with people who may suffer from a disorder affecting their quality of life, such an error is usually more serious than the error of over-referring (Ostergard, 1983:226).

It has already been established that the pure-tone audiogram is unable to detect differences in speech recognition in noise across listeners (Killion and Niquette, 2000:50), and is therefore not a sensitive measure of an individual's ability to understand speech in noise. Traditional word recognition tests have also been reported to be insensitive to differences in recognition performance across listeners with various degrees of hearing loss (Lucks Mendel and Danhauer, 1997:205). It has been noted in the literature that tests using sentence material as stimuli yield a higher precision in determining speech reception thresholds and is more sensitive to changes in speech recognition performance as test conditions change than tests using word material (Kollmeier and Wesselkamp, 1997:2415). This has been demonstrated by the fact that sentence materials yield a much steeper intelligibility slope than single words (Kollmeier and Wesselkamp, 1997:2415; Plomp and Mimpen, 1979:49). It is therefore important in the development of such a test that the intelligibility slope of individual sentences and lists be considered when selecting sentences for the final lists, as the efficiency of the test depends on this (Versfeld et al., 2000:1672,1673,1676).

The sensitivity and specificity of tests for sentence recognition in noise are not routinely reported by researchers developing such tests. However, Nilsson et al. (1994:1091) investigated the sensitivity of the American HINT by calculating the statistical power of the test to detect small differences in threshold. This was computed according to the standard deviations of differences between repeated measures. It was found that the sentence recognition task was more sensitive to detect threshold differences when presented in noise than under quiet conditions. It was also reported that using

a greater number of lists to determine a mean threshold had greater sensitivity in predicting thresholds than using a single list.

The sensitivity and specificity of a test can also be indicated by its ability to separate affected individuals from those with normal function in terms of the skill being assessed. Wilson, McArdle and Smith (2007) investigated four different speech-in-noise tests (the BKB-SIN, HINT, QuickSIN and WIN) in terms of their ability to separate hearing-impaired individuals from normal-hearing subjects. This was done by comparing the test scores of the normal-hearing individuals on each test with those of the hearing-impaired subjects. Findings indicated that the QuickSIN (Quick Speech-In-Noise test) and WIN (Words In Noise test) showed the greatest difference between normal hearers and those with a hearing impairment, indicating that these two measures may be more sensitive than the BKB-SIN and HINT (Wilson, McArdle and Smith, 2007:855).

However, they also found that the BKB-SIN and HINT materials were easier and yielded higher scores in both groups of subjects due to the greater amount of semantic cues in these materials (Wilson, McArdle and Smith, 2007:855; 846). This attribute could make these tests more useful in populations where poorer performance is expected, such as cochlear implant candidates or the paediatric population. It could also be said that these measures will then be more specific in these populations than the more difficult QuickSIN or WIN tests, since a greater number of these individuals will perform well on the easier tests, which would lead to less referrals.

In conclusion, the sensitivity and specificity of a developed measure are important indicators of test performance and should be considered during or after the development of such a test to ensure optimal performance. Due to the reciprocal relationship between these two variables, researchers should take into account the aim of the developed test as well as the target population to negotiate the ideal balance between these two variables.

2.8 Conclusion

The ability to understand speech under noisy conditions is an essential communication skill and therefore constitutes an important area to assess in audiology. Consequently, the development of tests to assess this function should enjoy priority within the field of speech audiometry, especially within developing contexts where little work has been done in this area. However, the perception of speech in the presence of noise is a complex process affected by many factors, and the development of tests assessing this ability should pay careful attention to all these aspects.

These aspects or variables can be arranged into four distinct categories, which together constitute a framework that can be used to guide the development or evaluation of a speech audiometry test. These categories include stimulus variables, presentation variables, subject variables, response variables and test performance variables. The existing literature provide both the basic theoretical underpinnings necessary to make decisions about these variables, as well as examples of previous studies aimed at the development of a test for speech perception in noise. Through a critical review of both the basic theory as well as previous studies of a similar nature, the current chapter provided a theoretical structure that clearly outlined the options available when developing a test of speech perception, specifically focusing on tests using sentence material in the presence of noise.

Table 2.7 provides a summary of the test method variables discussed in the previous sections. These variables are each outlined in terms of the different options reported in the literature, along with the relevant references. The purpose of this table is to provide a clear and concise outline to guide researchers in the development of new measures in speech audiometry.

Table 2.7: Summary of different test method variables

	TEST VARIABLE	OPTIONS	REFERENCES
STIMULUS	Composition of speech material	Develop own/original material	Plomp and Mimpen (1979), Vaillancourt et al. (2005), Versfeld et al. (2000), Van Wieringen and Wouters (2006), Cameron and Dillon (2007a)
		Adaptation of existing material	Nilsson et al. (1994), Wong and Soli (2005), Kollmeier and Wesselkamp (1997), Hällgren et al. (2006)
	Equalising sentence difficulty	Re-scale intensity of sentences that are too hard / too easy	Plomp and Mimpen (1979), Nilsson et al (1994), Hällgren et al. (2006), Wong and Soli (2005), Wong et al. (2007)
		Eliminating / excluding sentences that are too hard / too easy	Kollmeier and Wesselkamp (1997), Versfeld et al (2000), Vaillancourt et al (2005), Van Wieringen and Wouters (2006)
		Select subset or decide on re-scaling based on SNR-50 only	Plomp and Mimpen (1979), Nilsson et al. (1994), Wong and Soli (2005), Wong et al. (2007)
		Select subset or decide on re-scaling based on SNR-50 and psychometric slope	Kollmeier and Wesselkamp (1997), Versfeld et al. (2000), Vaillancourt et al. (2005), Hällgren et al. (2006), Van Wieringen and Wouters (2006)
	Noise type	Recorded speech (continuous discourse or multi-talker babble)	Kalikow et al. (1977), Cameron and Dillon (2007a)
Idealised speech weighted noise consistent with the mean LTASS spectrum across languages		Byrne et al. (1994), Wagener and Brand (2005)	
Noise created according to LTASS of recording specific to test being developed		Plomp and Mimpen (1979), Nilsson et al. (1994), Hällgren et al. (2006), Vaillancourt et al. (2005), Wong and Soli (2005), Van Wieringen and Wouters (2006)	
Gender of speaker	Male	Nilsson et al. (1994), Kollmeier and Wesselkamp (1997), Vaillancourt et al. (2005)	
	Female	Plomp and Mimpen (1979), Hällgren et al. (2006), Cameron and Dillon (2007a)	
	Male and female speaker	Versfeld et al. (2000), Van Wieringen and Wouters (2006)	
Training of speaker	Speech therapist / audiologist	Hällgren (2006), Van Wieringen and Wouters (2006), Wong et al. (2007), Cameron and Dillon (2007a)	

Table 2.7: Summary of different test method variables (continued)

PRESENTATION	Presentation method	Fixed presentation level in initial phases	Plomp and Mimpen (1979), Nilsson et al. (1994), Vaillancourt et al. (2005), Wong and Soli (2005), Van Wieringen & Wouters (2006), Wong et al. (2007)	
		Fixed presentation level throughout	Kollmeier and Wesselkamp (1997)	
		Adaptive presentation method once lists have been compiled	Plomp and Mimpen (1979), Nilsson et al. (1994), Vaillancourt et al. (2005), Wong and Soli (2005), Van Wieringen & Wouters (2006), Wong et al. (2007)	
		Adaptive presentation method throughout	Versfeld et al. (2000), Hällgren et al. (2006), Cameron and Dillon (2007a)	
PRESENTATION	Presentation level (speech and noise pre-mixed and re-scaled)	70 dB SPL	Versfeld et al. (2000), Wagener and Brand (2005)	
		Auditory transmission channel	Sound-field	Hällgren et al. (2006)
			Headphones (binaural)	Plomp and Mimpen (1979), Nilsson et al. (1994)
			Headphones (monaural)	Plomp and Mimpen (1979), Kollmeier and Wesselkamp (1997), Versfeld et al. (2000), Van Wieringen and Wouters (2006)
Headphones simulating noise front and side conditions	Wong and Soli (2005), Vaillancourt et al. (2005), Wong et al. (2007), Cameron and Dillon (2007a)			
SUBJECT	Subject characteristics	Auditory characteristics: Thresholds ≤ 15 dB HL from 250 to 8000 Hz, normal otoscopic examination, tympanograms, otologic history, no history/symptoms of auditory processing disorder	Nilsson et al. (1994), Versfeld et al. (2000), Vaillancourt et al. (2005), Hällgren et al. (2006), Neijenhuis et al. (2001), Crandell (1991), Bellis (2003b)	
		Age range: 18-30 years	Hällgren et al. (2006)	
		Language: Native speaker, Grade 12 education in test language	Vaillancourt et al. (2005)	
		Cognition: Completed Grade 12 in mainstream education	McLauchlin, (1980), Vaillancourt et al. (2005)	
RESPONSE	Response channel	Written / typed	Versfeld et al. (2000)	
		Verbal	Plomp and Mimpen (1979), Nilsson et al. (1994), Kollmeier and Wesselkamp (1997), Versfeld et al.(2000), Vaillancourt et al. (2005), Wong & Soli (2005), Hällgren et al. (2006), Van Wieringen and Wouters (2006)	
	Scoring method initial phase	Word-by-word	Plomp and Mimpen (1979), Nilsson et al. (1994), Kollmeier and Wesselkamp (1997), Versfeld et al.(2000), Vaillancourt et al. (2005), Wong & Soli (2005), Hällgren et al. (2006), Van Wieringen and Wouters (2006)	
Syllable-by-syllable		Not previously documented		

Table 2.7: Summary of different test method variables (continued)

TEST PERFORMANCE	VALIDITY	Face validity	High degree of apparent validity for estimating an individual's ability to deal with typical speech stimuli	Lucks Mendel and Danhauer (1997)
		Content validity	Test material considered natural by native speakers of the test language Use noise as part of the test material to make test condition more representative of everyday listening environment	Nilsson et al. (1994), Vaillancourt et al. (2005), Hällgren et al. (2006), Wong and Soli (2000) Hällgren et al. (2006)
		Criterion-related validity	Compare test results with existing tests	Wilson, McArdle and Smith (2007)
			If no existing tests to compare with, refer to construct validity	Ostergard (1983)
	Construct validity	Apply test to hearing-impaired population with deficit in evaluated skill Simulate hearing deficit in normal-hearing population	Van Wieringen and Wouters (2006) Stuart et al. (1995), Scott et al. (2001)	
	RELIABILITY	Stability	Test-retest reliability evaluated by re-testing subjects after a period of time	Hällgren et al. (2006)
		Equivalence / Inter-list reliability	Achieved by phonetically balancing lists	Plomp & Mimpfen (1979), Nilsson et al. (1994), Hällgren et al. (2006), Vaillancourt et al. (2005), Wong and Soli (2005) Nilsson et al. (1994), Wong and Soli (2005), Vaillancourt et al. (2006), Hällgren et al. (2006), Wong et al. (2007) Kollmeier and Wesselkamp (1997), Versfeld et al. (2000)
			Achieved by arranging lists according to intelligibility slopes	
	Assess by comparing mean score for each list across subjects with overall mean			
	Internal consistency	Assess according to standard deviation of list means across subjects		
Compare thresholds obtained by a certain combination of lists with thresholds from a different combination of lists Compare threshold obtained by one list to those obtained with other lists by an analysis of variance		Vaillancourt et al. (2005)		
SENSITIVITY	Sensitivity / specificity	Determine statistical power to predict a true difference in threshold	Nilsson et al. (1994)	
		Apply test to both normal-hearing and hearing-impaired individuals and assess ability of test to separate these groups	Wilson, McArdle and Smith (2007)	

2.9 Summary

This chapter reviewed the existing literature and provided the theoretical underpinnings required for the development of a test for sentence recognition in noise by exploring the variables involved in such a test, including the stimulus, presentation, subject, response and performance variables. A model of these variables was provided and a critical discussion of the relevant literature followed. Conclusions were drawn from the literature review and the different variables and options for each were subsequently summarised in table format.

3. METHODOLOGY

3.1 Introduction

The preceding chapter provided a theoretical framework and critical review of existing literature that served to guide the methodology of the study, which will be described in this chapter. The study consisted of three distinct phases, each with its own aim, participants, and procedures. The present chapter describes the methodology by specifying the aims, research design, ethical considerations, research sample, material and apparatus, procedures, as well as the validity and reliability of the research. A summary of the procedures of each phase is also provided in table format at the end of the chapter.

3.2 Aims

The main aim of the study was to develop a valid and reliable Afrikaans test of sentence recognition thresholds in noise.

In order to attain this main aim, the project consisted of three phases, each with a separate aim, as stipulated below.

Sub-aim 1: To develop a collection of recorded Afrikaans sentences suitable for the assessment of speech recognition in noise (Phase I).

Sub-aim 2: To select, from the recorded material, a collection of sentences with equivalent intelligibility in the presence of noise (Phase II).

Sub-aim 3: To compare inter-list reliability and response variability of two list sets compiled using two different methods of list compilation (Phase III).

3.3 Research Design

The overall design of the study can be described as design-based. Although this paradigm is mostly applied to educational contexts (The Design-Based Research Collective, 2003:5), it provides a suitable framework for the current study. In the conventional application of this method, theoretical expertise is used to design a particular learning environment or intervention, which is then applied in an educational context. The context itself is seen as a core part of the process, and not just an extraneous variable (Barab and Squire, 2004:3). The process and results of the application are then used in a series of processes where flexible design revision is used to systematically improve initial designs (Barab and Squire, 2004:2, 3; Wang and Hannafin, 2005:6).

The principles of this approach were applied to the present study by using theoretical knowledge to compile test material potentially suitable for the assessment of sentence recognition in background noise. This material was subsequently applied to a number of subjects under different conditions. Each application yielded a set of results needed to implement the next project phase, but also provided valuable findings that contributed to the current body of knowledge. In this way, the results of applying the test material served to systematically improve the developed measure and enabled suitable revision of the application. This process is shown in Figure 3.1.

The input of Phase I (indicated in the figure by the thick blue arrows) consisted of speech material from various sources. The cyclic process in Phase I (indicated by the grey arrows) entailed the adaptation of this material to suit the goals of the current research, as well as the rating of grammar level and naturalness. This whole process resulted in two sets of data, namely basic and applied data. In the diagram, the purple arrow and text box indicate the basic data. This data set encompassed the findings that were not essential for the conduction of the following phase of the project, but yielded valuable guidelines that were used to refine the research process, and could also be used in future studies of a similar nature.

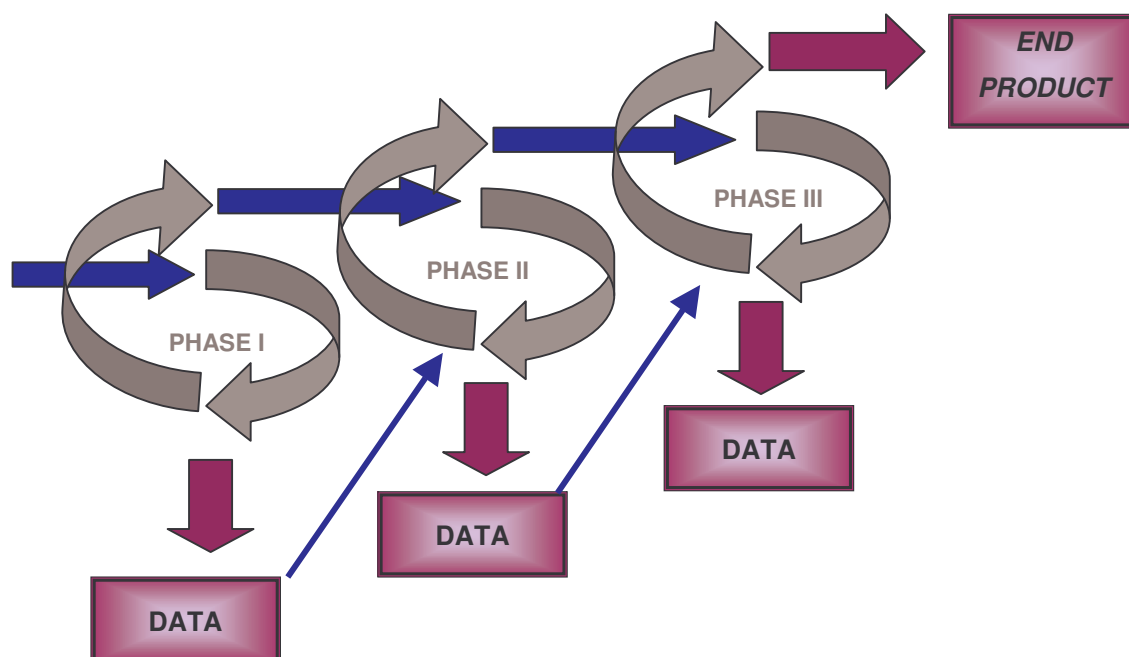


Figure 3.1: Design-based research cycle as applied to the current study

The applied data that resulted from the first phase constituted the input (again indicated by the blue arrow) of the following phase. This input consisted of the sentence material that had been refined through the processes of the first phase. The smaller blue arrow pointing from the “Data” text box to the following phase refers to the basic data that were not essential to the next phase, but nevertheless contributed to the quality of the process. In the case of the first phase, the results of the grammar rating yielded such data, since this was used to conduct additional analyses on the findings from Phase II.

A similar process was followed during Phases II and III, with each phase yielding both basic and applied data. The final phase resulted in an end product, namely an Afrikaans test for speech recognition in noise. In addition, the final phase yielded valuable data on the methodology for compiling equivalent lists for such a test. This data was used to refine the final product, but also provided valuable information that could assist future researchers in the process of developing a similar test. All the data collected during the current research was quantitative or numerical in nature (Leedy and Ormrod,

2005:102). The details of the design and methods followed in each of the three phases are provided in the procedure sections of this chapter.

3.4 Ethical considerations

As long as human participants are used in research, their treatment during, and benefit from the research process should remain an important consideration (Salkind, 2006:58). The ethical aspects described below were therefore taken into account in the planning of this research project.

3.4.1 Respect for the privacy of research participants

In order to ensure the privacy and confidentiality of all the participants (Huysamen, 1994:183; Hegde, 2003:546), no individuals were named in the research report. Where necessary, research subjects were designated a specific code for data processing and the names of participants were not used in data analysis or reporting. This was clearly explained in an informed consent letter provided to all participants.

3.4.2 Informed consent

According to Hegde (2003:545), informed consent is a crucial ethical principle and entails three components – the participants should fully comprehend the research procedure, they should give free and willing consent for participation and should be free to withdraw from the research at any point in time. Subsequently, the present study acquired written informed consent by having the appropriate form (attached as Appendix A) signed by each participant, after reading a letter explaining the goals and procedures of the study (see Appendix B). The letter also stated clearly that participation is voluntary and participants may withdraw from the study at any time and assured participants of confidentiality.

3.4.3 Beneficence and non-maleficance

In order to attain the second and third sub-aims of the project, volunteers underwent an evaluation of speech recognition. During this evaluation, sound intensities were kept at a comfortable level (around 70 dB SPL) and were therefore not harmful to the subjects (Salkind, 2006:58). The potential inconvenience of participating in the study (the time and effort required to undergo the test) was indicated in the letter of consent. The letter also explained that the information gathered should provide useful data to speech, language and hearing professionals in South Africa by publication of results upon conclusion of the study (Hegde, 2003:544). The participants in the second and third phases benefited from the process in that they received a free hearing screening test during selection procedures. No other incentives or rewards were offered for participation in the study.

3.4.4 Distributive justice

The effect of potential inequality between the researcher and research participants was limited by the fact that all subjects were of the same cultural/linguistic background as the researcher. In addition, the majority of participants (Groups B to G as listed in Table 3.1) were roughly the same age as the researcher.

3.4.5 Cultural and linguistic diversity

One of the primary selection criteria was that participants be native speakers of Afrikaans (i.e. must consider Afrikaans their home language). Therefore, cultural and linguistic diversity as an ethical issue did not have any effect on this particular project.

3.5 Research Sample

The current research project involved several groups of participants that were involved in different phases of the project. These participants all belonged to

the population described below, but adhered to different selection criteria, that will be specified for each group separately.

3.5.1 Population

The total population from which the participants for the study were selected was native speakers of Afrikaans in South Africa. The total population of South Africa in 2007 was reported to be around 47.9 million (Statistics South Africa, 2007:1), of which approximately 13.3% are reported to regard Afrikaans as their home language (Statistics South Africa, 2001:5).

3.5.2 Selection criteria

The participants of this research project were all selected from the total population of Afrikaans-speaking South Africans. Six distinct groups of participants were selected to participate in the different processes involved in the research project, each according to specific selection criteria. The process and phase that each of these groups were involved in are summarised in Table 3.1.

Table 3.1: Summary of participant groups

GROUP	PROCESS	PHASE	SAMPLING TECHNIQUE
Group A	Rating of naturalness	Phase I	Purposive sampling
Group B	Potential speakers for recording	Phase I	Convenience sampling
Group C	Speaker selected for recordings	Phase I	Purposive sampling
Group D	First equalisation procedure (selection of equivalent subset of sentences)	Phase II	Convenience sampling
Group E	Second equalisation procedure (determining intelligibility slopes)	Phase II	Convenience sampling
Group F	Experimental application of "slope lists" (compiled according to intelligibility slopes)	Phase III	Convenience sampling
Group G	Experimental application of phonetically balanced lists	Phase III	Convenience sampling

As shown in Table 3.1, two different sampling techniques were used for selecting the participants. For Groups A and C purposive sampling (Leedy and Ormrod, 2005:206) was used. This sampling technique entailed purposefully selecting specific individuals according to pre-determined criteria. Groups B, D, E, F, and G were selected using convenience sampling (Maxwell and Satake, 2006:96), as the sample consisted mainly of students and other young adults living in and around Pretoria that were available and willing to participate. The researcher aimed to reduce the limitations of convenience sampling by specifying certain selection criteria as illustrated in Table 3.2 below.

Table 3.2: Selection criteria for each group of participants

GROUP	CRITERIA	REFERENCE
A	Representative of different age groups, occupations, geographical and educational backgrounds	Vaillancourt et al. (2005)
B	Female speakers, aged between 18 and 30 years	
C	Standard dialect or absence of dialectical influences Clear articulation Suitable voice quality and intonation Appropriate loudness and speech rate Pronunciation not reflective of obvious personal or social characteristics Pronunciation should not be breathy, untidy, dialectical, or conceited	Vaillancourt et al. (2005); Cameron and Dillon (2007a) Versfeld et al. (2000); De Villiers and Ponelis (1987)
D, E, F, G	Hearing thresholds \leq 15 dB HL at 250, 500, 1000, 2000, 4000 and 8000 Hz Normal otoscopic examination, normal tympanograms, negative otologic history Age range: 18 – 30 years Completion of Grade 12 in mainstream education No history of neurologic disease (tumour or stroke), neurosurgery, Traumatic Brain Injury, childhood APD, or clinically significant difficulty to hear speech in noise No previous exposure to sentence material	Nilsson et al. (1994), Versfeld et al. (2000), Vaillancourt et al. (2005), Hällgren et al. (2006), Wong et al. (2007) McLauchlin (1980) Bellis (2003b)

The table shows that Groups A, B, and C each had distinct selection criteria. Groups D to G constituted the listeners used in the experimental application of

the recorded material, and were all selected according to the same criteria. It should also be noted that in addition to the criteria stipulated in the table, all the participants in each group were required to be native Afrikaans speakers, that is, Afrikaans should be the language that they speak at home most frequently (Statistics South Africa, 2001:5). Group C of the subjects (the speaker selected for recordings) was selected from Group B according to the selection criteria stipulated in Table 3.2.

3.5.3 Selection procedures

The participants in each group were selected according to specific procedures to ensure their adherence to the selection criteria. These procedures will subsequently be described as applicable to each specific group.

3.5.3.1 Group A

This research sample was purposefully selected to represent a wide variety of occupations, educational backgrounds, geographical origins and age groups. This was confirmed by the completion of a form questioning the potential participant on each of these aspects (attached as Appendix C).

3.5.3.2 Group B

It was decided to use a female speaker for the recordings. Therefore, the potential speakers for the recordings were required to be Afrikaans-speaking females between the ages of 18 and 30. To select this group, students and lecturers in the Department of Communication Pathology at the University of Pretoria that adhered to these criteria were informed of the research project and the individuals that volunteered to participate were selected to be included in this group.

3.5.3.3 Group C

The speaker selected for the recording of the sentence material was selected from the group of potential speakers described as Group B. The selection

process entailed recording a speech sample from each of the speakers in Group B, and submitting these recordings to a panel of judges that consisted of two speech therapists and two audiologists. The judges listened to the recordings and completed a rating scale describing the intelligibility, naturalness, articulation, speech rate, vocal quality, resonance, intonation, dialect and overall impression of each speaker. This was completed on a form such as the one attached in Appendix D. The ratings given by the judges were quantified as illustrated in Table 3.3 below.

Table 3.3: Rating scale for speakers

Intelligibility	Good = 3	Ave = 2	Poor = 1
Naturalness	Good = 3	Ave = 2	Poor = 1
Articulation	Good = 3	Ave = 2	Poor = 1
Voice quality	Good = 3	Ave = 2	Poor = 1
Resonance	Good = 3	Ave = 2	Poor = 1
Intonation	Good = 3	Ave = 2	Poor = 1
Speech rate	Correct = 1	Too fast = 0	Too slow = 0
Affected / overarticulated	No = 1	Yes = 0	
Accented / dialectical	No = 1	Yes = 0	
General impression	/10		

As shown in the table, judges had to allocate a rating of good, average or poor for intelligibility, naturalness, articulation, voice quality, resonance and intonation. This three-point scale was quantified so that a “good” rating received 3 points, an “average” rating received 2 points, and a “poor” rating received 1 point. Speech rate was judged as either appropriate for such material, too fast or too slow. Both the “too fast” and “too slow” ratings were quantified as a score of 0, since both these ratings would make a speaker equally unsuitable for the recordings. An appropriate speech rate was scored as 1. Overarticulation and accented speech were both judged to be either present or absent – a speaker that was judged to overarticulate or present with a specific regional dialect or accent received scores of 0 for these aspects. Finally, judges had to allocate a score out of ten to each speaker to summarise their overall impression of the speaker’s suitability to the task.

After quantifying the ratings, the mean rating for each aspect of each speaker according to all four judges was calculated. Subsequently, the means of all the aspects (for each speaker separately) were added to obtain a total score for each speaker. In this manner, a total score out of 30 was calculated for each of the participating speakers. The speaker with the highest score (28.5) was selected for the recordings.

3.5.3.4 Groups D to G

These groups of participants volunteered for participation in the study, and those adhering to the stipulated selection criteria were identified by a series of selection procedures. The following selection procedures were carried out for each potential participant.

1. Completion of a case history form (included in Appendix E) by the researching audiologist. Questions covered the subjects' age, otologic history, educational history (Grade 12 in mainstream Afrikaans school), language most often spoken at home, as well as risk factors for Auditory Processing Disorders (Vaillancourt et al., 2005:360; McLauchlin, 1980:253; Bellis, 2003b:10).
2. An otoscopic examination was done by gently pulling back the pinna and placing the speculum of an otoscope in the ear canal opening. Visual inspection of the ear canal and tympanic membrane was conducted to identify any abnormalities such as excessive cerumen or any signs of ear disease (Roush, 2001:36). These results were interpreted in conjunction with the results of the tympanometry.
3. Screening tympanometry was conducted with the use of a hand-held screening tympanometer. A clean probe tip was used for each participant and the tympanometer automatically displayed the result on its screen once a seal was obtained in the ear canal. A Type A tympanogram (single peak with base-peak compliance difference in the range of 0.3-1.6cc occurring at 0 mm \pm 50mm) was considered to

indicate normal middle ear functioning (Wiley and Fowler, 1997:53; Hannley, 1986:55).

4. Screening audiometry to ensure that audiometric thresholds were 15 dB HL or better at octave frequencies from 250 to 8000 Hz. The procedures for screening were based on the guidelines provided by the American Speech-Language-Hearing Association (2005:4-5). Following instructions to the subject to say “yes” every time a tone is heard, however soft it might be, testing commenced with the presentation of a 1000 Hz tone at 30 dB HL. After a clear response, intensity was decreased by 10 dB, and increased with 5 dB after each failure to respond. This adaptive up-down procedure was followed until a clear response was confirmed (at least two out of three positive responses at that intensity) at a level below 20 dB HL (i.e. if a subject showed a clear, confirmed response at 15 dB HL, the audiologist proceeded to the following frequency).

3.5.4 Description of research sample

Groups A, B and C participated in Phase I of the project, Groups D and E in Phase II and Groups F and G in Phase III. A description of each of these groups is provided in this section.

3.5.4.1 Group A

This group constituted the subjects that provided a rating of naturalness for the sentence material and consisted of 10 individuals, 5 of whom participated in the first round of naturalness rating, and 5 in the second round. These subjects all adhered to the selection criteria stipulated in Table 3.2, and a description of the sample is provided in Table 3.4 below. The table indicates the gender, age, occupation, educational and geographical background of each subject, as all of these factors were considered during the selection of this participant group.

Table 3.4: Description of subjects in Group A

NO.:	SEX	AGE	CURRENT OCCUPATION	HIGHEST ACADEMIC QUALIFICATION	AREAS WHERE EDUCATION WAS RECEIVED
1	F	53	Receptionist	Tertiary Diploma in Higher Education	Pretoria, Gauteng
2	F	29	Speech therapist	B Communication Pathology	Pretoria, Gauteng
3	M	40	Used car salesman	Trade test in aircraft maintenance mechanics	Bloemfontein and Edinburgh, Free State
4	F	21	Hairdresser	Diploma in Hair Technology	Carltonville, North West Province
5	M	85	Pensioner	Advanced Language Examination (Afrikaans and English)	Rustenburg, North West Province Technikon Pretoria, Gauteng
6	F	34	Orders clerk	Tertiary Diploma in Textile Design	Kempton Park, Gauteng
7	M	55	Management consultant	B.Sc (Honours)	Vanwyksvlei, Northern Cape Cape Town, Western Cape
8	M	22	Student	B.Com (Human Resources)	Delmas, Mpumalanga Howick, Kwazulu Natal
9	F	48	Cabin attendant	Grade 12	Potgietersrus, Limpopo Warmbaths, Limpopo
10	M	95	Pensioner	BA, M.Ed.	Bronkhorstfontein & Potchefstroom, North West Province

The information provided in this table indicates that the subjects' ages ranged from 21 to 95 years, with a mean age of 48. Furthermore, within the total number of subjects, a great variety of backgrounds occurred, representing ten different occupations and eight of South Africa's nine provinces.

3.5.4.2 Group B

The second group of subjects (Group B) consisted of the potential speakers that volunteered for the recording of the material. This group consisted of eight females, all native speakers of Afrikaans, with ages ranging from 18 to 28, and a mean age of 20 years.

3.5.4.3 Group C

Group C of the participants consisted of a single female speaker that was selected from Group B. This speaker was a 26-year old speech and language therapist that was judged to have good articulation, intelligibility, voice quality, and resonance as well as an appropriate speech rate and intonation, without a specific accent or over-articulated, unnatural speech quality.

3.5.4.4 Groups D to G

Groups D and E participated in the second phase of the project. The first group (Group D, $n = 10$) consisted of 5 males and 5 females, and participated in the first equalisation procedure, while the second group (Group E, $n = 12$) consisted of 6 males and 6 females who took part in the second procedure. Groups F and G took part in the third phase of the research. Group F was called the experimental group and Group G the control group. These last two groups consisted of 10 subjects each (5 male, 5 female), and subjects were randomly assigned to either group. In total, 22 individuals (11 male, 11 female) participated in Phase II, and 20 in Phase III (10 male, 10 female). Table 3.5 below indicates the age of each of the participants in the order that they were tested.

Table 3.5: Ages of participants in Groups D, E, F, and G

PHASE II				PHASE III			
GROUP D	AGE	GROUP E	AGE	GROUP F	AGE	GROUP G	AGE
1	29	1	23	1	26	1	26
2	21	2	24	2	20	2	28
3	24	3	23	3	20	3	22
4	27	4	22	4	25	4	23
5	18	5	27	5	28	5	22
6	23	6	25	6	22	6	20
7	27	7	19	7	19	7	23
8	24	8	24	8	18	8	19
9	23	9	23	9	23	9	22
10	26	10	24	10	30	10	19
		11	23				
		12	21				
Mean age:	24		23		23		24
Std dev:	3.22		1.99		4.04		2.88

3.6 Material and apparatus

The material and apparatus used during the research project can be divided into material used for the selection of research subjects, and that used during data collection, analysis and recording. Both these categories will be dealt with in the following section.

3.6.1 Material and apparatus for subject selection

Table 3.6 provides an overview of the material and apparatus used for the selection of each group of research subjects, along with the processes that these groups were involved in.

Table 3.6: Material and apparatus used in selection of each group of subjects

GROUP	PROCESS	MATERIAL / APPARATUS
Group A	Rating of naturalness	“Profile of participant (Group A)” form (Appendix C)
Group B	Potential speakers for recording	-
Group C	Speaker selected for recordings	Fostex M-2 Unidirectional handheld microphone Nakamichi 550 Versatile Cassette System Maico Speakers Sound-treated laboratory “Rating of speakers” form (Appendix D)
Group D, E, F, G	Listeners in experiments, Phases II and III	“Case history form (Groups D-G)” (Appendix E) Welch Allyn otoscope Interacoustics MT10 screening tympanometer Madsen audiometer with Standard TDH39 headphones, calibrated for use with the audiometer Double-walled sound-proof booth (maximum ambient sound pressure level exceeds specifications as set out in SABS 0182; highest ambient sound level was measured at 4.7 dB SPL at 125 Hz, and lower levels at higher frequencies; calibrated on 7 August 2006)

The “Profile of participant (Group A)” as well as “Case history form (Groups D-G)” listed in the table were compiled to ascertain the adherence of potential participants to the selection criteria specified in Table 3.2.

3.6.2 Material and apparatus for data collection, recording and analysis

The material and apparatus used during the various phases of the project for data collection, recording and analysis are shown in Table 3.7. This table provides a clear overview of the equipment and material used during the research, along with the purpose or process where this material was applied and the phase in which each particular process occurred. It should be noted that the recorded speech material is not described in this table, since the development of this material is described throughout the report. The format and content of the speech material also changed throughout the research and the final collection of sentences is described in the results section of the study.

Table 3.7: Material and apparatus used during various data collection, recording and analysis procedures

MATERIAL / APPARATUS	PURPOSE DURING RESEARCH	PHASE
BKB sentences (Bench and Bamford, 1979) “Afrikaanse Reseptiewe Woordeskattoets”, Test Form A (Buitendag, 1994) Afrikaans word discrimination lists for children aged 3-5, Tesner and Laubscher, unpublished)	Resources for compilation of sentence material	Phase I
Letter accompanying informed consent form (Appendix B)	Explained research process to subjects	Phase I, II & III
Informed consent form (Appendix A)	Signed by all research subjects	Phase I, II & III
Instructions for rating of naturalness (Appendix F)	Explained rating procedure to subjects	Phase I
Rating form (naturalness)	Used by first 5 subjects in Group A to record their rating and comments on naturalness	Phase I
Second rating form (naturalness)	Used by subjects 6-10 in Group A to record their rating of sentences	Phase I
Sennheiser ME62 microphone	Recording of speech material	Phase I
Creative Labs Soundblaster Extigy external sound card (sampled at 44.1 kHz with 24-bit resolution)	Converting signal from analogue to digital during recording of speech material	Phase I
Creative Wave Studio software, version 4.21.04 ⁴	Interface for recording of speech material	Phase I
Test form 1	Recording in subjects' responses during presentation of material at fixed SNRs	Phase II
Test form 2 (Appendix G)	Test form containing “slope lists”, used to record SNR-50 thresholds obtained during adaptive test procedure	Phase III
Test form 3 (Appendix H)	Test form containing phonetically balanced lists, used to record SNR-50 thresholds during adaptive test procedure	Phase III

⁴ Available from Creative Labs (<http://us.creative.com/>)

Table 3.7: Material and apparatus used during various data collection, recording and analysis procedures (continued)

Double-walled sound-proof booth (highest ambient sound level measured at 4.7 dB SPL at 125 Hz, and lower levels at higher frequencies; calibrated on 7 August 2006)	Used as recording and testing environment	Phase I, II & III
“Praat” software: Doing phonetics by computer, version 4.4.24 (Boersma & Weenink, 2006)	Used to edit recorded sentences	Phase I
	Used as interface to present sentences at fixed SNRs	Phase II
Matlab: Software for doing mathematics on a computer, version 7.0.1 ⁵	Used to generate a speech-weighted noise, with spectral envelope matching the average power spectral density of the entire set of 515 recorded sentences	Phase II
	Used to calculate phonetic content and errors in phonetic balance during compilation of phonetically balanced lists	Phase III
	Used as interface for adaptive presentation method to determine SNR-50 thresholds	Phase III
Madsen Midimate 622 diagnostic audiometer with standard TDH39 headphones	Used to present the pre-recorded material	Phase II & III
Pentium IV personal computer with a 2.4GHz hard drive	Used to record, store and present sentence material in .wav format	Phase I, II & III
M-Audio Fast Track Pro mobile USB Audio Interface with preamplifiers (samples at 96kHz with 16-bit resolution; signal-to-noise ratio for line outputs on the preamplifier specified to be -103dBA ⁶)	Used to route the recorded material from the computer to the auxiliary input of the audiometer during testing with both fixed and adaptive SNR methods	Phase II & III
BMDP Statistical Software, version 7.1 ⁷	Used for statistical analysis of results	Phase II & III

⁵ Available from The MathWorks Inc. (www.mathworks.com)

⁶ As specified by M-Audio (http://www.m-audio.com/products/en_us/FastTrackPro-focus.html)

⁷ Available from BMDP Statistical Software Inc.

As shown in the table, a form was supplied to the participants in Group A to record their rating of the naturalness of the sentence material. This form was completed by the first 5 subjects in Group A. It contained a total of 518 sentences which had to be rated for naturalness on a scale of 1 to 7. The rating form provided space for a rating (1-7) as well as for suggested changes to sentences that received a rating lower than 6. The second rating form listed in the table contained only the sentences that were altered according to the recommendations of the first 5 subjects. On the second round of rating, subjects were only required to provide a rating from 1 to 7 for each sentence, and did not have to provide suggestions for change, since sentences that received a low rating during this round were rejected from the collection.

During Phase II of the project, the recorded sentence material with added speech-weighted noise was presented to a total of 22 subjects. Responses were recorded on a form (Test form 1 in Table 3.7) that specified the number and content of each sentence, as well as a column for comments, a blank column for the subject's syllable score (number of syllables repeated correctly) and a column specifying the total number of syllables of that particular sentence.

The test forms used during Phase III of the research project (Test forms 2 and 3 in Table 3.7) differed from the test form used in Phase II in that the sentences were now arranged into lists of 10 sentences each and the form served as a template to judge the correctness of subjects' responses and to record the SNR-50 threshold obtained for each list. These test forms are included in Appendices G and H.

3.7 Procedures: Phase I

The first phase of this project represented the first cycle in the design-based research process illustrated in Figure 3.1. The aim of this phase was the development and recording of a large collection of sentences potentially

suitable for a sentence recognition test. Information necessary to refine and describe the sentence material was collected using quantitative surveys. This method of data collection can be called non-experimental, descriptive research (Leedy and Ormrod, 2005:183). The processes of this phase are described in this section.

3.7.1 Compilation of sentences

An examination of the literature showed that there are no formally standardised sentence tests or collections of suitable sentences available in Afrikaans. Therefore, a large collection of sentences was compiled using two methods, namely translation of existing material and compilation of original material similar in content and structure to the translated material. These two methods were combined to enlarge the sentence collection, since a larger collection would allow for the elimination of sentences that were found through experimental evaluation to be unsuitable for the final collection. Details of these eliminations will be provided under the methodology of subsequent phases. Throughout the process of translation and compilation of additional sentences, the objective was to compile sentences that are complete, (containing at least one verb), representative of everyday speech, and free from proverbs, questions, exclamations and proper nouns, as these characteristics ensure that sentences are not too redundant or confusing (Versfeld et al., 2000:1672; Plomp and Mimpen, 1979:44).

The material selected for translation consisted of a large set of short sentences that were designed for use with British children, namely the BKB or Bamford-Kowal-Bench sentences (Bench and Bamford, 1979). The verbs and nouns contained in the material were found in transcriptions of British children's speech. These sentences were selected due to the size of the collection as well as its simplicity in terms of syntax and grammar. Furthermore, these sentences were used as a starting point for the American HINT test material and were found suitable for this use (Nilsson et al., 1994:1086). It should be noted that although the present study was not aimed

at developing a test for children, using this type of material ensures that sentences are all of approximately the same known level of complexity. Furthermore, since individuals with hearing impairment often have limited language abilities (McLauchlin, 1980:253), material used to evaluate their hearing should be kept as simple as possible.

The translation of the material was done by a speech language therapist whose first language is Afrikaans and who had lived in the United Kingdom for two years, thereby ensuring knowledge of both cultures. This was important to ensure that the translated material would be free from expressions that are not culturally relevant in Afrikaans. The translator also had two years' part time experience in simultaneous translation of Afrikaans into English. The translation procedure encompassed the following four steps.

1. A survey of the literature was conducted to find previous translations of the material (Pakendorf, 1998:5). One such translation was found (Cloete, 1997), but the content of the material was loosely translated, and was never evaluated for naturalness or its use in background noise. Therefore, it was decided to re-do the translations according to criteria suitable for the aims of the current research.
2. The BKB sentences were translated into Afrikaans, initially focusing mainly on naturalness rather than accuracy of translation since the goal was to compile sentences that would sound natural in Afrikaans, and the test would not evaluate the listener's knowledge of the specific content or syntax but rather his/her ability to decipher typical Afrikaans sentences in the presence of background noise.
3. The translated material was revised to ensure that the content of the sentences had been accurately translated.
4. The translations were revised a second time to enhance the uniformity of the material by ensuring that the sentences were all reasonably similar in length (with the number of words between 4 and 8, and the number of syllables between 4 and 9). In order to achieve this, it was sometimes necessary to change the tense of the sentence. Additional

changes had to be made to the content of a number of sentences in order to enhance their naturalness in Afrikaans. The type of changes made to the content are summarised and motivated in Table 3.8 below.

Table 3.8: Changes made to sentence content

TYPE OF CHANGE	MOTIVATION
Semantical	Direct translation would result in sentences that are too long or would sound artificial/unnatural
Tense / time	Direct translation retaining the same tense would make sentences too long
Cultural	Direct translation would contain vocabulary or concepts that are not commonly used by South Africans

As shown in the table, three types of changes were made. Semantical changes entailed changing one or more words in the sentence, thereby altering the meaning and improving the naturalness. Changes to the tense in which the sentence was written were applied where the original tense would have resulted in an inappropriately long or awkward-sounding sentence. Finally, “cultural” changes meant that typical British vocabulary was replaced with words or concepts more appropriate to the South African culture.

Following the translation of the BKB sentences, additional sentence material similar in structure and content was composed in order to enlarge the sentence collection. This was achieved by using a collection of words known to be comprehensible for young children as a basis for the creation of new sentences (Vaillancourt et al., 2005:360). This method was used to ensure that the keywords in the sentences were of equal, known difficulty. Two collections of words considered to be suitable for evaluating young children were chosen for the present study.

The first collection was the “Afrikaanse Reseptiewe Woordeskat toets” (Afrikaans Receptive Vocabulary Test) or ARW and was developed by

Buitendag in 1994. It consists of two separate lists of words, equivalent in difficulty and showing a high reliability and validity for the evaluation of vocabulary of Afrikaans-speaking children aged two to twelve years (Buitendag, 1994:154). Each list is divided into different age sections, thereby giving an indication of the age level at which the listed words may be understood by children. These characteristics made this test a valuable source of vocabulary suitable for a speech-in-noise test.

The second source of vocabulary was the so-called “Phonetically Balanced” Word Lists for children aged three to five. These lists were compiled by Tesner and Laubscher in 1966, although the methodology was never published. The words in these lists were mostly taken from children’s books and were mainly selected on account of their familiarity to children. According to Tesner (personal communication, September 4, 2006), these lists, despite its name, are not phonetically balanced or representative, as the main focus in its compilation was familiarity of the vocabulary to young children. Although these lists have not been formally standardised in any published form, they have been used for the evaluation of word discrimination in young children with great success for the past forty years and were therefore considered a suitable resource for the compilation of sentence material.

The words taken from the ARW and the “Phonetically Balanced” children’s word lists (PBC lists) were as keywords to compile new sentences. These sentences were created to be similar in length and structure to the translated BKB sentences. This was accomplished by ensuring that these sentences each had a number of syllables between four and nine.

3.7.2 Rating of naturalness

After compilation, the sentence material was submitted to native speakers of Afrikaans in order to ensure that the developed material was considered acceptable and natural by the general population (Nilsson et al., 1994:1086; Hällgren et al., 2006:228; Wong and Soli, 2000:278). Subjects taking part in

this rating process were specifically selected to differ in age, occupation and geographical background in an attempt to ensure a research sample representative of a large part of the South African population, as suggested by the method of Vaillancourt et al. (2005:360).

Initially, the entire collection of sentences was submitted to five participants (2 male and 3 female). A description of these subjects is provided in Table 3.4 (participant numbers one to five). After informed consent was obtained, participants were supplied with a written copy of the sentences as well as written instructions (Appendix F). The instructions directed participants to rate the naturalness of each sentence on a scale from 1 to 7, where 7 would indicate that the sentences sounded like a natural (in other words realistic) Afrikaans sentences, and 1 would mean that the sentence sounded completely artificial. Furthermore, participants were instructed to provide suggestions for a more natural sentence construction for each sentence rated lower than 6 (Nilsson et al., 1994:1086). Two written examples were also provided to illustrate the rating procedure.

Subsequently, sentences that were altered after the first rating were submitted to a second panel of participants (three male, two female). These participants also represented various occupations, age groups, and backgrounds (see Table 3.4, participants six to ten). During the second rating, participants were required to rate only the sentences altered after the first rating, on the same 7-point scale described for the first round of rating. They did not have to provide suggested alternatives, as sentences that received a poor rating in this round were not revised but rather rejected from the collection.

3.7.3 Rating of grammatical complexity

After the naturalness rating, sentences were submitted for analysis of grammatical complexity by an expert in language development and speech analysis. This was done to evaluate the uniformity of the sentences in terms of grammar. The expert was a speech and language therapist with a PhD in

Communication Pathology, who has presented numerous lectures and seminars on language development and analysis.

Due to the fact that the sentences would eventually be presented without any meaningful context supporting their comprehension, they could only be rated in terms of grammatical structure and not in terms of vocabulary or semantical content. Also, rating sentences in this manner does not necessarily provide an indication of the receptive language level required to understand them, but rather provides a system according to which items may be grouped in terms of difficulty (E. Naudé, personal communication, January 22, 2007).

The following steps were followed in the analysis.

1. Sentences were analysed in clauses and phrases.
2. The age level of clauses was recorded according to the profile for the development of grammatical structures in Afrikaans (Naudé, 1998:96-98).
3. Phrase structures that occur after the age of 3 were noted and recorded.
4. The level of complexity was rated as the age level of clauses with an additional “+” for every phrase structure that appears after the age of 3.

Consequently, sentences were classified to be on one of the following levels of grammatical complexity.

1. Age 3 with no complex phrases
2. Age 3 with one complex phrase
3. Age 3 with two complex phrases
4. Age 4 with no complex phrases
5. Age 4 with one complex structure (clause or phrase)
6. Age 4 with two complex structures

7. Age 5 and more complex

3.7.4 Recording and editing of material

After a suitable speaker was selected, sentences were recorded digitally in a sound-proof booth. The microphone was placed on a microphone stand approximately 20cm from the speaker's mouth. Sentences were recorded one-by-one, and the speaker aimed to clearly articulate all words, without distortion of any sounds and attempted to place equal emphasis on all parts of the sentence and maintain vocal effort throughout each sentence (Versfeld et al., 2000:1672), while still retaining a natural intonation pattern.

The speaker in the current recordings aimed for a standard pronunciation by avoiding the abnormal and striving for the general form as far as possible (Le Roux and Pienaar, 1976:xviii). This is because there are many varieties (standard and non-standard) of Afrikaans (Carstens, 2003:282), and non-standard varieties such as regional dialects are usually characterised by differences in pronunciation, sentence structure, semantics and peculiarities in vocabulary (Carstens, 2003:289). In contrast, a standard variety is defined as the form of this language that serves as a model against which users of the language measure the way they talk or the variety accepted by the language community as the language form with the greatest value for use across the entire language territory (Carstens, 2003:283, 285).

The spectrogram of each sentence was visually inspected after recording to look for overemphasis of any sounds or words (indicated by increased amounts of energy on the waveform). Each recorded sentence was also played back through a loudspeaker, and if any mistakes or distortions were identified, the sentence was re-recorded. All the recordings were done within a period of three days, in sessions of 45 minutes each, interrupted by breaks of at least 15 minutes to prevent vocal fatigue from influencing the quality of the recordings.

After completion of the recordings, waveforms were edited by eliminating unwanted silences preceding and following the recorded speech. Maximum silent intervals of 0.08 to 0.1 ms were allowed. Subsequently, the average intensity of each sentence was checked, and it was found that all sentences had an average intensity of 65 to 77 dB. The intensity of each sentence was then re-scaled to 70 dB before saving it to hard disc in .wav format.

A random sample of 15 sentences was then submitted to an expert to assess whether the pronunciation of the recorded speaker could be considered as standard Afrikaans. The expert had a Master's degree in speech science and forty years' experience lecturing in phonetics at a tertiary institution. The pronunciation was considered to be of a standard variety and was therefore deemed appropriate for the goals of the current project.

3.8 Procedures: Phase II

The procedures and results of Phase II represented the second cycle of the design-based research process (see Figure 3.1). The input for this cycle was the collection of 515 digitally recorded sentences that resulted from the first phase. The desired output of the second phase was a sentence collection that would be more uniform in terms of intelligibility. This output was produced using a non-experimental descriptive design (Leedy and Ormrod, 2005:179). The acquisition of quantitative data on the sentence material enabled the researcher to refine and reduce the sentence collection.

In order to produce the desired output of this phase, two procedures aimed at the equalisation of the sentence collection were followed. It should be noted that the equalisation was attained by means of the elimination of sentences that did not fall within specified criteria, and not by re-scaling intensities. The first equalisation procedure served to improve the uniformity of the collection by eliminating sentences that did not yield performances within one standard

deviation of the mean performance at one fixed SNR. This procedure also reduced the size of the sentence collection, which in turn reduced the testing time required during the second equalisation procedure, when only the sentences remaining after the first procedure were presented to the research participants. During this second procedure, the intelligibility slopes of these sentences were determined and sentences presenting with a slope that did not fall within specified criteria were eliminated. This entire process is illustrated in Figure 3.2, which shows that the first equalisation procedure yielded data necessary for the completion of the second procedure. The second procedure resulted in an output that was essential for the conduction of the third phase of the project, namely a subset of sentences known to be of equivalent difficulty and to have a similar intelligibility slope. The details of the procedures followed during each of these two equalisation procedures are described in this section.

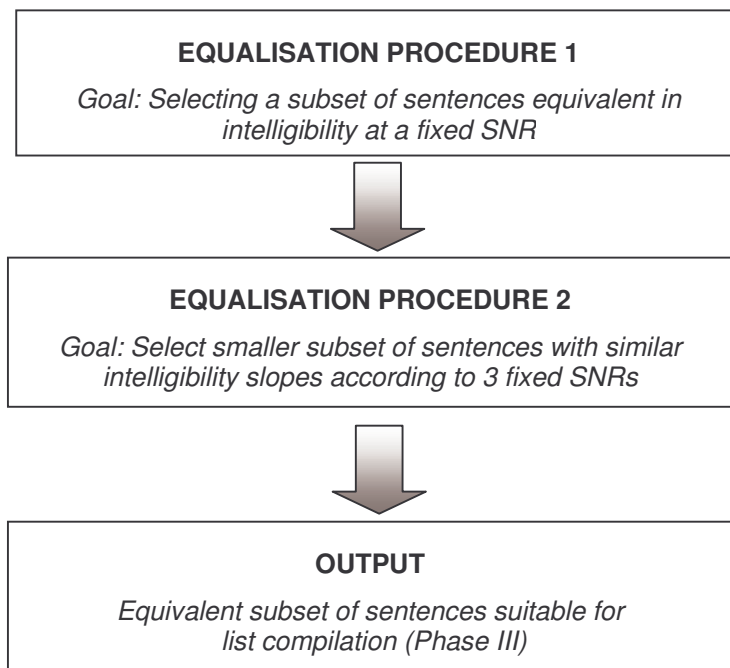


Figure 3.2: Research design of the second phase

3.8.1 First equalisation procedure: Selection of equivalent subset of sentences

The first procedure in this phase was aimed at increasing the equivalence of the sentence material and reducing the initial number of sentences by eliminating those sentences that were significantly easier or harder than the majority of sentences. This was achieved by presenting the entire collection of sentences to ten different subjects in the presence of a fixed amount of background noise.

For these purposes, a suitable background noise had to be created and added to the recorded speech material. Speech-weighted noise with a spectral envelope matching the average power spectral density of the entire set of 515 recorded sentences (Nilsson et al., 1994:1087; Hällgren et al., 2006:228) was generated in Matlab. The sentences with the added noise were each saved as a separate wave file, numbered according to their position in the list.

For this first equalisation procedure, the noise was added to each sentence to ensure an SNR of -5 dB (noise intensity 5 dB larger than speech intensity). This level was chosen in accordance with the findings of previous researchers (refer to Table 2.4, Chapter 2), which indicated that an SNR of -5 dB should yield an intelligibility score (defined in this instance as the percentage of syllables correctly discerned) of approximately 50%. A 50% score was selected as a target score in an attempt to avoid the ceiling or floor effects that a higher or lower score might have on the procedure, especially since the difficulty level of the sentences was largely unknown at this stage.

3.8.1.1 Data collection

Each of the ten subjects in this group was tested separately. The subject was seated in a sound-proof booth with the test administrator (a qualified audiologist). The total collection of 515 sentences was presented to each subject at a pre-mixed SNR of -5 dB, scaled to an intensity of 70 dB SPL. The

sentences with the added noise were presented to both the left and right ear simultaneously, using headphones as transducer. Each sentence was presented once only and frequent breaks were given in between to prevent exhaustion of the subject. The order of presentation of the sentences was counterbalanced by arranging sentences into 10 “playlists” before testing, and starting each subject on a different “playlist”. The composition of the playlists is shown in Table 3.9.

Table 3.9: Composition of playlists for 1st equalisation procedure

Playlist 1	Sentence 1-50
Playlist 2	Sentence 51-100
Playlist 3	Sentence 101-150
Playlist 4	Sentence 151-200
Playlist 5	Sentence 201-250
Playlist 6	Sentence 251-300
Playlist 7	Sentence 301-350
Playlist 8	Sentence 351-405
Playlist 9	Sentence 406-460
Playlist 10	Sentence 461-515

As shown in Table 3.9, each playlist contained 50 sentences, except for the last 3 lists, which each contained 55 sentences. The goal of counterbalancing the presentation order of the lists was to mitigate the effect that practice might have on performance. It was anticipated that performance might improve somewhat within the first list presented as the subject adapted to the task. For this reason, each subject started with a different list.

Prior to testing, all subjects were provided with the following information regarding the test procedure.

- They would be listening to Afrikaans sentences uttered by a female speaker.
- The sentences were all short and simple, and without obscure or unpredictable content.

- ❑ The sentences did not contain any proper nouns, idioms, or questions.
- ❑ There would be noise added to the speech, which would be louder than the sentences and might make the listening task quite difficult.
- ❑ Each sentence would only be played once.

Subsequently, subjects were instructed to repeat back aloud what they had heard every time, even if it was only part of a word or sentence, and were encouraged to guess at the content if uncertain.

3.8.1.2 Data recording and analysis

During the test procedure, the administrator used a test form to record the responses of the subjects. The form contained a text version of each of the 515 sentences and the total number of syllables of each sentence. The collected results (number of correct syllables for each sentence) were subsequently entered into an electronic spreadsheet, which was used to calculate the percentage of syllables repeated correctly for each sentence.

The mean percentage score for each sentence across subjects was calculated, as well as the mean score for each subject across the total collection of sentences. In order to proceed to the second equalisation procedure with a smaller, more uniform collection of sentences, the mean percentage score as well as the standard deviation for all sentences across all subjects was determined. Sentences that fell within one standard deviation of the mean were included for testing in the second procedure.

In addition to the calculations necessary to identify the sentences for the next procedure, a number of other procedures were also conducted to further analyse the data collected in the first procedure. This was done with the help of the Department of Statistics at the University of Pretoria. During data collection, it was noted that there appeared to be a difference between the performance of male and female participants, and an analysis of variance (ANOVA) procedure was performed on the data to determine the significance

of this difference. Furthermore, the grammar rating assigned to each sentence in Phase I was also compared to the performance scores obtained in the present phase to determine the correlation between the intelligibility of a sentence and its grammatical complexity. This was also done using an analysis of variance.

Furthermore, the effect of a sentence's position in the presentation order was investigated to explore the effect of practice on performance. This was achieved by re-arranging the results of each subject according to the order in which the sentences were presented and subsequently comparing the scores of the first 10 sentences as well as the score for the 11th to 20th sentences. A Friedman two-way analysis of variance (ANOVA) was used to determine whether differences between these scores were significant.

3.8.2 Second equalisation procedure: Selecting sentences with similar intelligibility slopes

The second equalisation procedure in this phase was aimed at determining the intelligibility slope (percentage score as a function of SNR) for each sentence, in order to select sentences with similar slopes for the following phase. This slope was determined by plotting the intelligibility score of each sentence at three different SNRs (SNR-5 from the previous procedure, as well as the two SNR conditions tested in this second procedure). The subset of sentences selected after the first equalisation procedure ($n=330$) were used in this second procedure.

3.8.2.1 Data collection

The same procedures described for data collection of the first equalisation procedure (under 3.8.1) were followed for the second procedure, with the exception of the SNR at which the material was presented. Two different SNRs were used in this procedure. The first six subjects of Group E (3 males and 3 females) listened to the 330 sentences at an SNR of -8 dB (i.e. the noise was 8 dB louder than the speech), while the last six subjects (3 male, 3

female) listened to the recordings at an SNR of -2 dB. The noise was pre-mixed with the signal to attain the required SNR, and the combined signal was then re-scaled to 70 dB SPL.

Due to an apparent learning effect observed in the first procedure, a list of practice sentences was presented to each subject before commencing the test. Practice lists were compiled as follows:

- ❑ Twenty sentences were selected from the sentences rejected after the first selection procedure (SNR-5). Ten of these sentences were rejected because they were too difficult, and ten because they were too easy. The percentage scores of the twenty selected sentences did not differ more than 2% from the upper and lower limits of the range of sentences selected for testing, ensuring that their level of difficulty would not differ considerably from the subsequent sentences.
- ❑ These 20 sentences were arranged into 2 playlists consisting of 10 sentences each, with 5 easy sentences first in each list, followed by 5 difficult sentences.

Despite this additional measure to prevent a practice effect, the order of presentation of the different playlists was still counterbalanced so that each subject heard a different playlist first. The 330 sentences were thus divided into 6 playlists of 55 sentences each and presented in the order illustrated in Table 3.10 below.

Table 3.10: Playlist order for second equalisation procedure

SUBJECT 1	SUBJECT 2	SUBJECT 3	SUBJECT 4	SUBJECT 5	SUBJECT 6
Practice lists	Practice lists	Practice lists	Practice lists	Practice lists	Practice lists
List 1	List 2	List 3	List 4	List 5	List 6
List 2	List 3	List 4	List 5	List 6	List 1
List 3	List 4	List 5	List 6	List 1	List 2
List 4	List 5	List 6	List 1	List 2	List 3
List 5	List 6	List 1	List 2	List 3	List 4
List 6	List 1	List 2	List 3	List 4	List 5

3.8.2.2 Data recording and analysis

The same type of test form used for the first procedure (displaying the correct version of the sentence, the total number of syllables, as well as room for comments and scores) was used to record the results. However, the test form for this second procedure contained only the 330 sentences selected during the first equalisation procedure. The recorded results were subsequently entered into an electronic spreadsheet and analysed with the assistance of the Department of Statistics at the University of Pretoria. The goal of the data analysis was to select from the 330 sentences used in the second procedure a subset of sentences with similar speech intelligibility slopes that would yield similar results at specific SNRs. This was achieved as follows.

- ❑ The mean percentage score (syllables correct) for each sentence across listeners at each specific SNR (-2, -5 and -8) was calculated.
- ❑ The mean score and standard deviation for the total collection of 330 sentences at each specific SNR was calculated.
- ❑ The differences between scores at SNR-8 and SNR-2 (Difference 3-1), SNR-8 and SNR-5 (Difference 2-1), and SNR-5 and SNR-2 (Difference 3-2) were calculated for each sentence separately. This process is illustrated in Figure 3.3.
- ❑ These differences were also calculated across the entire collection of sentences as a group (mean differences and standard deviations).

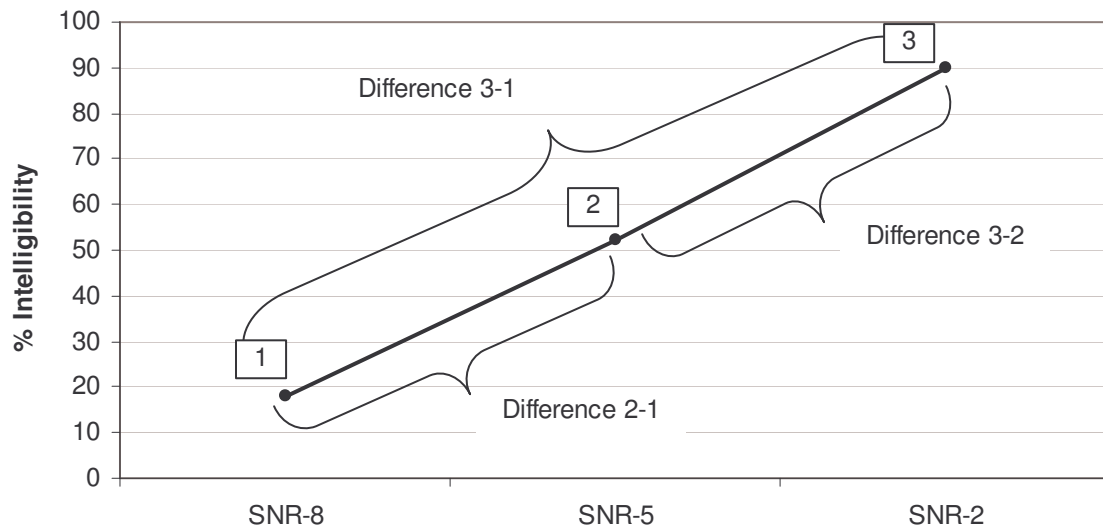


Figure 3.3: Mean intelligibility slope, illustrating the method followed in selecting sentences with a similar slope

- Subsequently, sentences were selected that adhered to the following criteria.
 - The difference between the percentage score at SNR-8 and SNR-2 (Difference 3-1 as illustrated in Figure 3.3) had to fall within one standard deviation of the mean difference across all sentences for these two points on the slope. This implied that the mean slope of the sentence between the two extremes had to fall within a certain area.
 - The difference between the percentage score at SNR-8 and SNR-5 (Difference 2-1) as well as the difference between the score at SNR-5 and SNR-2 (Difference 3-2) had to fall within two standard deviations of the mean difference of these points across all sentences. This ensured that between each of the extremes and the centre of the slope, each sentence was required to have a certain slope. By implication, this meant that no sentences that had a suitable mean slope between the two extremes could have a negative slope or extremely flat/steep slope between any two points on the graph.

3.9 Procedures: Phase III

The third phase of the project constituted the final cycle in the design-based research process described in Chapter 3. The data input for this cycle was the 222 sentences selected during the second phase, and the output was the final selection of sentences, arranged into lists of equivalent difficulty. The additional data output of this phase concerned the method of list compilation, which provided valuable guidelines for future research of a similar kind.

The research methodology followed to complete this cycle can be described as experimental. In this type of research the experimenter manipulates certain variables, introduces one or more controls over the experiment and randomly assigns subjects to a control or experimental group (Polit and Hungler, 1991:152). All the aspects controlled by the researcher (the test method, list content and subject characteristics in this case) can be described as independent variables. The reliability or inter-list variability and resulting SNR-50 values were the dependent variables.

The procedures of the third phase were divided into three major processes, namely:

1. Compilation of lists.
2. Experiment I – application of lists to normal-hearing group.
3. Experiment II – application of lists to listeners with simulated hearing loss.

Two different methods of list compilation were followed. The first method has not been previously documented, whereas several other researchers have employed the second method. Therefore, the first set of lists (called “slope lists”) was the experimental method. The second method entailed compiling a set of lists that were balanced in terms of phonetic content. These lists were called “phonetically balanced lists” (PB lists) and represented the control method.

The second and third processes involved the evaluation of the inter-list equivalence or reliability of these two sets of lists. This was achieved with the use of two distinct experiments, as illustrated in Figure 3.4.

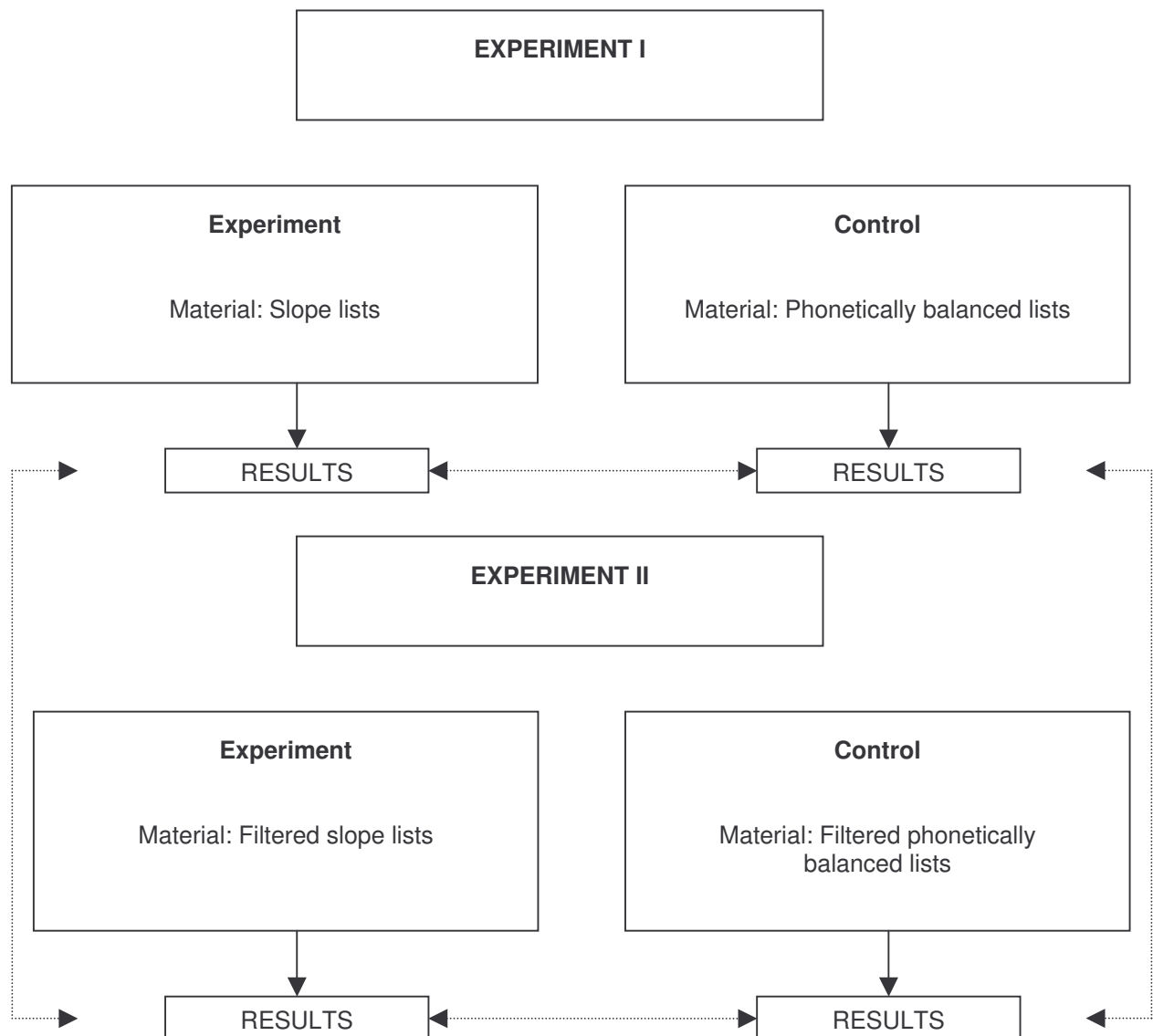


Figure 3.4: Research process of Phase III

The first experiment, as depicted in the figure, served to assess the reliability of the lists in a group of normal-hearing listeners, whereas the second experiment explored the effect that a simulated hearing loss had on this

reliability. The hearing loss was simulated by filtering the material used in the first experiment, as indicated in Figure 3.4 under Experiment II. Results from the two experiments as well as results from the experimental and control conditions could then be compared, as indicated by the dotted lines in Figure 3.4. The procedures followed during list compilation and subsequent experiments are discussed in this section.

3.9.1 Compilation of lists

An important requirement for speech material used in the determination of thresholds levels is that the materials used in different trials should be of equal, known difficulty. Furthermore, speech material should be different for each trial since it becomes easier if re-used or repeated (Nilsson et al., 1994). These requirements imply that a test using sentence material should have a collection of different sentence lists that are of equivalent difficulty. A well-reported method of ensuring list equivalence is to arrange sentences into phonetically balanced lists (Plomp and Mimpen, 1979:45; Nilsson et al., 1994:1088; Hällgren et al., 2006:229; Vaillancourt et al., 2005:362, Wong and Soli, 2005:282). However, in light of the questions in the literature regarding the value and necessity of phonetically balancing lists (Tobias, 1964:99; Carhart, 1970:229; Martin et al., 2000:489), a method of list compilation mainly focusing on list equivalence, and not phonetic content as such, was explored in the present study. Since this was an experimental method, the well-reported method of phonetically balancing lists was also conducted as a control measure.

3.9.1.1 Slope lists

The experimental method involved grouping sentences together based solely on intelligibility slopes i.e. the percentage scores obtained at different SNRs. In this method, the first 220 sentences were initially grouped into lists of ten in numerical order. The mean performance score for a list of ten sentences at each SNR previously tested (in Phase II) was calculated and plotted as a slope. These slopes are illustrated in Figure 3.5 below.

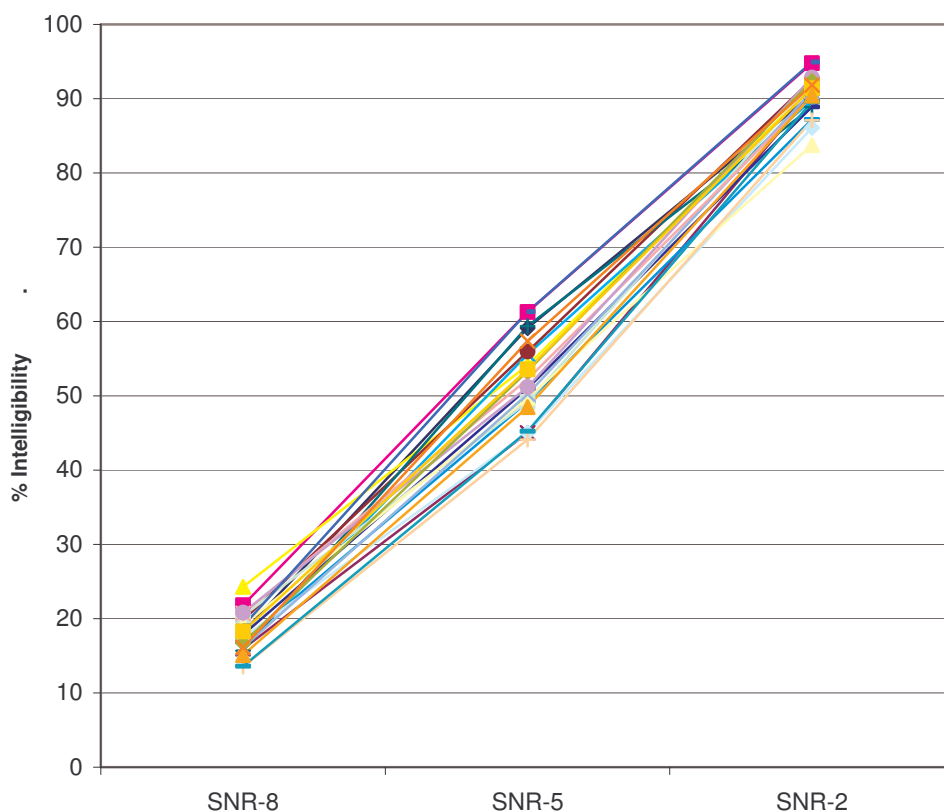


Figure 3.5: Intelligibility slopes for lists grouped according to numerical order

The difference between the minimum and maximum values at SNR-8, SNR-5, and SNR-2 were 10, 17 and 11% respectively. Lists with points on the slope that deviated most from the other lists were then edited by exchanging sentences from those lists with sentences in other lists. For example, if List 1 showed a very poor performance mean at SNR-8, the sentence in that list with the lowest score at that SNR was exchanged with a sentence in another list that had a relatively high score at that particular SNR. Likewise, a list showing a performance point too high at a certain data point (SNR) was edited by exchanging a sentence with a high score at that point with a sentence from another list that had a low score for that point. The lists compiled in this manner were called “slope lists”, since they were compiled according to their intelligibility slopes.

3.9.1.2 Phonetically balanced lists

In order to be able to evaluate the validity and efficiency of this new method of list compilation and compare it with a well-documented method, the more traditional method of compiling phonetically balanced lists was also followed as a control procedure. This was achieved by arranging the same 222 sentences in different lists, this time according to phonetic content. In order to achieve this phonetic equivalence, phoneme occurrences had to be determined for each list, and sentences interchanged between lists in order to balance phoneme occurrence across lists (Plomp and Mimpen, 1979:45). Previous researchers have achieved this with the use of a trial-and-error exchange process, followed by the application of these lists to test subjects in order to evaluate its equivalence (Nilsson et al., 1994:1088).

In the current study, the 222 sentences selected through experimentation and statistical analyses in Phase II were submitted to an expert in speech analysis for phonetic transcription. Transcription was initially done from the written material, but then reviewed according to the digitally recorded material. Since the process of balancing phonemic distribution across lists was to be conducted with the help of an electronic spreadsheet, transcription was done using symbols that were either available on a standard keyboard or could be inserted using a keyboard shortcut. The symbols used in the present study, along with the traditional phonetic symbols (Carstens, 2003:297-299) for each phoneme are shown in Table 3.11.

Table 3.11: Phonetic symbols used for transcription in spreadsheet

PHONETIC SYMBOL		NEW SYMBOL	PHONETIC SYMBOL		NEW SYMBOL
Vowels	[a]	a	Consonants	[b]	b
	[ɑ:]	ɑ:		[d]	d
	[e]	e:		[f]	f
	[ɛ]	ê		[g]	g
	[ɛ:]	ê:		[x]	x
	[æ]	ae		[ç]	ç
	[æ:]	ae:		[k]	k
	[ə]	e		[c]	c
	[i]	l		[l]	L
	[i:]	l:		[m]	m
	[o]	o:		[n]	n
	[ɔ]	o		[ŋ]	ng
	[ɔ:]	ô		[ɲ]	jn
	[∅]	eu		[p]	p
	[œ]	oe		[r]	r
	[œ:]	oe:		[s]	s
	[u]	u		[ʃ]	sj
[u:]	u:	[t]	t		
[y]	y	[tʃ]	tj		
[y:]	y:	[v]	v		
Nasalised vowels	[ã:]	ã	[w]	w	
	[ɛ̃:]	ë	[h]	h	
	[]	ö	[j]	j	
	[œ̃:]	ü	[z]	z	
	[ẽ:]	ï	[ʒ]	dz	
Diphthongs	[ai]	ai	[ʔ]	1	
	[ɑ:i]	ɑ:l			
	[oi]	o:l			
	[ɔ:]	oi			
	[ui]	ui			
	[əi]	ei			
	[eu]	eeu			
	[œu]	oeu			
[œy]/ [œi]	oey				

An electronic spreadsheet was used to calculate the total number of occurrences of each phoneme in the collection. These occurrences are indicated in Table 3.12 below, with the phonemes arranged in order of descending occurrence (from most frequent to least frequent) in each group (vowels, diphthongs, and consonants separately).

Table 3.12: Phoneme occurrences arranged according to frequency of occurrence

PHONEME		OCCURRENCE	PHONEME		OCCURRENCE
Vowels	e	356	Consonants	t	300
	l	253		s	295
	a	131		r	253
	a:	123		d	242
	ê	108		L	195
	o	101		h	189
	e:	62		k	174
	o:	60		n	172
	u	44		l	164
	oe	41		x	137
	ae	30		p	129
	eu	15		m	113
	y:	8		b	109
	ö	8		f	107
	ê:	5		v	46
	ô	5		c	20
	u:	5		ng	19
	y	4		w	8
	l:	3		j	7
	ä	1		jn	6
ë	1	g	0		
ae:	0	ç	0		
oe:	0	sj	0		
ü	0	tj	0		
ï	0	z	0		
Diphthongs	ei	138	dz	0	
	oey	34			
	oeu	19			
	ai	9			
	o:l	6			
	a:l	4			
	ui	2			
	oi	1			
	eeu	0			

The total number of occurrences for each phoneme as shown in the table was then divided by the total number of sentences (222) to determine the average occurrence of each phoneme per sentence. As sentences would be arranged into lists of 10, the average occurrence per sentence was multiplied by ten to determine the number of times that each phoneme should occur in a list of 10 sentences. This was then considered the “ideal” number of times that this phoneme should occur in each list.

Since the aim was to compile 22 lists from the 222 sentences, it followed that a phoneme occurring approximately 22 times would have to occur once per list. In order to begin list compilation, sentences were therefore first sorted according to the occurrence of a phoneme that occurred 22 times or as close as possible to (but not less than) 22. The closest phoneme to this was “ae” (see Table 3.12), which occurred 30 times in the entire collection. Sentences that contained this phoneme were distributed across the 22 lists (one sentence per list).

The number of occurrences of each of the phonemes in each of these lists (although currently only consisting of one sentence) was calculated and compared to the ideal number of times this phoneme should occur per list. If a phoneme deviated from this ideal number, it was considered an error. For instance, the “a” phoneme was supposed to occur 6 times per list (total occurrence of 131 divided by 22 and multiplied by 10 = 6). If a list contained 4 “a” phonemes (2 less than the ideal 6), it had an error value of 2 for this phoneme. This type of error (where there was a shortage of a certain phoneme) was called a positive error. An excess of a particular phoneme (for example 8 instead of the ideal 6) was considered a negative error. The total error value of a list could then be calculated as the sum of the errors on all the phonemes.

Subsequently, the 200 sentences that had not been assigned to a list yet, were substituted one by one in place of the sentences already assigned to lists. For each substitution, the total number of positive errors for the list was calculated. After all the unused sentences had been tried into each list, the sentence that yielded the smallest number of positive errors was chosen (one per list). If two sentences resulted in the same number of positive errors, the total positive error value of these sentences was calculated, and the sentence with the smallest total positive error was chosen.

Every list then had one sentence that was as close as possible to ideal in terms of its phonetic content. Subsequently, a second sentence was added to each list. This was done using the following procedures.

- ❑ Each of the unassigned (remaining) sentences was added to a particular list one by one and the total error for the list was calculated.
- ❑ The sentence that yielded the lowest error value in a particular list was assigned to that list.
- ❑ If the same sentence yielded the lowest error value for more than one list, it was assigned to the list in which it gave the lowest total error.
- ❑ Once each list had two sentences, the remaining (unassigned) sentences were each tried as a substitution for each of these two sentences. If a “new” sentence yielded a smaller number of positive errors for a particular list, the substitution was made.

This same process was repeated until each list contained 10 sentences and all the lists had the lowest possible error. During this process of substitution, the negative error on each phoneme was constantly monitored to ensure that it was not greater than 2 for any one phoneme. This meant that whenever a sentence was added to a list and it yielded a negative error larger than 2 (i.e. an excess larger than 2) for any phoneme, it was immediately rejected. Once no more changes could be made that would reduce the errors, and lists still contained less than 10 sentences, this negative error limit was increased to 3 per phoneme and the remaining sentences were added to the lists.

This process resulted in a total of 22 phonetically balanced lists consisting of 10 sentences each. The details of these lists will be provided in Chapter 4 (section 4.4.1.2). The compiled lists were then subjected to an experimental evaluation, which will be described in the following section.

3.9.2 Experimental application of lists

Following the compilation of the two sets of lists, both the experimental (slope) lists and control (PB) lists were subjected to experimentation to evaluate its reliability. Two separate experiments were conducted, each with its own control group. The first experiment evaluated the reliability of the lists in a group of normal-hearing listeners, whereas the second experiment aimed to test the reliability of the lists in a group of listeners with a simulated hearing loss. For both experiments, the slope lists constituted the experimental condition, and the PB lists served as control. The results of the experimental and control conditions could then be compared, as well as the results of the two experiments.

3.9.2.1 *Experiment I*

The same test procedures and equipment were used for both the experimental and control conditions. The material and apparatus used during this phase are listed in Table 3.7 and copies of the test forms used during this experiment are attached as Appendices G and H. The research sample was discussed under section 3.5 of this chapter.

During experimentation, test subjects were seated in the sound-proof booth with the researcher. Prior to testing, all subjects were provided with the following information regarding the test procedure.

- They would be listening to Afrikaans sentences uttered by a female speaker.
- The sentences were all short and simple, and without obscure or unpredictable content.
- The sentences did not contain any proper nouns, idioms, or questions.
- There would be noise added to the speech, which would be louder than the sentences and might make the task quite difficult.
- Sentences were grouped into lists of ten.

- ❑ The first sentence in each list would be repeatedly played to them, until they were able to correctly repeat it.
- ❑ Subsequent sentences in each list would be played only once, regardless of the correctness of the repetition thereof.

Subjects were then instructed to repeat back aloud what they had heard every time, even if it was only part of a word or sentence, and were encouraged to guess at the content if uncertain.

The order of list presentation was counterbalanced between subjects, with the first subject starting with List number 1, the second with List number 4, and all subsequent subjects starting with the next even-numbered list. Testing was preceded with the presentation of two practice lists each consisting of 10 randomly selected sentences that were rejected in the final experiment of phase II. An adaptive up-down presentation method was followed (Nilsson et al., 1994:1089). The level of the speech signal was kept constant, while the noise level was altered to produce the desired SNR. The combined speech-and-noise signal was scaled to an intensity of 70 dB SPL. The first sentence in a list was presented at an SNR estimated to be below threshold (-8 dB, in this case). This sentence was repeatedly presented with SNR levels improving in steps of 2 dB until the subject was able to correctly repeat it. The accuracy of the listener's response was compared to a text version of the sentence. Once the first sentence was correctly repeated, the next sentence was presented at the same SNR. The SNR levels of subsequent sentences were determined each time by the correctness of the preceding sentence's repetition. If a sentence was repeated correctly, the following sentence was presented at a more difficult SNR (noise level increased by 2 dB with speech level kept constant). If a sentence was repeated incorrectly, the following sentence was presented at a better or easier SNR (noise level decreased by 2 dB).

Since the SNR level of each sentence was determined by the correctness of the previous sentence's repetition, swift decisions had to be made on the

accuracy of repetitions. To facilitate this rapid judgement, stricter scoring criteria than previous phases were used. For the experiments in this phase, all words in a sentence had to be repeated accurately in order for it to be considered correct. Although previous researchers have allowed for small variations on words such as “a” and “the” (Nilsson et al., 1994:1089), criteria for this study was kept as simple and consistent as possible by requiring a 100% correct repetition.

At the end of a list of 10 sentences, the software calculated the SNR-50 as an mean of the presentation levels of the fifth to eleventh sentences (although the eleventh sentence was never presented, its presentation level could be determined according to the correctness of the tenth sentence’s repetition). This resulting SNR value was then recorded on the test form.

3.9.2.2 Experiment II

Since the sentence recognition test developed during this project could potentially be used to assess hearing-impaired individuals, it was important to ensure that compiled lists are of equivalent difficulty to this population, and not just to listeners with normal hearing. The effect that a hearing loss has on an individual’s speech recognition depends partly on the frequency range affected by the hearing loss. This is because the high frequency components that characterise some of the consonants are lost (Davis, 1970:92). It could thus be said that the hearing loss acts as a type of “frequency filter”. If two different lists in a speech-in-noise test therefore contain different amounts of phonemes that are particularly difficult for a person with hearing impairment due to their frequency content, the results will differ and the lists are therefore not equivalent. Therefore, although a normal-hearing listener may find the two lists to be of equal difficulty, a listener with a hearing impairment may not perform similarly on two lists with different frequency characteristics.

To evaluate the effect of a high frequency hearing loss on the equivalence of the slope lists, this type of hearing loss was simulated. The simulation was

accomplished by creating a low-pass filter with a cut-off frequency of 2000 Hz, and a roll-off slope of 48 dB per octave (Stuart et al., 1995:660; Scott et al., 2001:439) and using this to filter the sentence materials. This method of simulation does not account for all the effects of a cochlear hearing loss, and the results obtained in this manner will therefore not be used to accurately predict the scores that may be obtained by hearing-impaired subjects. However, this type of simulation was considered sufficient to evaluate the effect that such a loss would have on list equivalence.

The filtered sentence materials were then presented to the same normal-hearing subjects that had previously listened to the unfiltered version. It was postulated that the filter should not affect the inter-list reliability of the PB lists, since the phonetic content, and thus the frequency content of these lists was equivalent. In addition, the method of phonetically balancing sentence lists has been well documented as a successful method of ensuring list equivalence (Plomp and Mimpen, 1979:45; Nilsson et al., 1994:1088; Hällgren et al., 2006:229; Vaillancourt et al., 2005:362, Wong and Soli, 2005:282). For these reasons, the filtered slope lists were considered the experiment, and the filtered PB lists were used as a control.

For both sets of lists, the same group of subjects was used in Experiment I (unfiltered condition) and Experiment II (filtered condition). The reason for using the same listeners, and for using normal-hearing listeners instead of hearing-impaired individuals, was to limit the number of variables introduced in these experiments. Using a different group of listeners, normal-hearing or hearing-impaired, would mean that there could be differences in age, memory and auditory processing between the two groups that could influence performance and contaminate results. In contrast, re-testing the same group of subjects meant that the only variable that was altered for the second experiment (filtered condition) was the frequency characteristics of the speech material.

Although re-testing the same subjects with the same lists could mean a somewhat improved performance on the second test due to the effect of memory and learning, it should be noted that it was not the absolute values in SNR attained in the second experiment that were of paramount importance, but rather the relative differences or variability between the different lists. Therefore, the same method of re-testing the same group of subjects with a filtered version of the lists was followed with both sets of lists in order to compare the inter-list equivalence of each set in the filtered and unfiltered conditions. A waiting period of one week was still allowed between testing in an attempt to reduce the effect of memory on the test results (Hällgren et al., 2006:231).

3.10 Reliability and Validity

The reliability and validity of a test for sentence recognition in noise has been reviewed in Chapter 2. However, a distinction should be made between the validity and reliability of the measure being developed and the validity and reliability of the research project aimed at developing this measure. This section will deal with the validity and reliability of the research project as a whole (as opposed to the validity and reliability of the developed measure, which has been discussed in Chapter 2).

Reliability refers mainly to the consistency or repeatability of a test (Ostergard, 1983:224). When conducting a research project, it is recommended that researchers address the issue of reliability by evaluating the test-retest, parallel forms, split-half and internal consistency reliability of the measure used to collect data during the research process (Struwig and Stead, 2001:130-131). However, in the present study, the measures used to collect data were not existing measures, since the aim of the study was the development of a specific measure. The considerations regarded to ensure the reliability of the developed measure have been discussed in Chapter 2 (under test performance variables) and will again be appraised when the results of the research are discussed (Chapter 5).

Validity pertains to the extent to which a research design is scientifically sound or appropriately conducted (Struwig and Stead, 2001:136), and includes internal and external validity (Leedy and Ormrod, 2005:97). Internal validity is related to the issue of whether independent variables, and not other extraneous variables, are responsible for variations in the dependent variable (Struwig and Stead, 2001:136). If extraneous variables are not controlled, it is impossible to know whether the changes in the dependent variable were due to the independent variable, extraneous variables, or both.

There were a number of different extraneous variables that had to be considered in the design of the study in order to reduce the influence of these variables on the results. The following variables were considered and controlled (Struwig and Stead, 2001:137; Ventry and Schiavetti, 1980:67-81).

❑ Maturation

This effect pertains to changes that may occur in scores or performance due to normal growth or development in the research sample (Ventry and Schiavetti, 1980:69). To eliminate the effect that maturation might have on the performance of subjects, selection criteria stipulated that subjects' ages had to fall within a certain range (18-30 years). If subjects are re-assessed after a period of time, this effect should also be considered. In the current research, the subjects in Phase III subjects were all re-tested. However, only a week was allowed in between the first and second tests to limit the possibility of any changes occurring in hearing or auditory processing.

❑ History

Occasionally, events unrelated to the independent variables may occur in between testing and re-testing and thereby influence findings. This applies particularly to studies where a pre-test is applied, followed by

some form of treatment, and a post-test is then conducted to evaluate the effects of the treatment. The current study did not include any treatment, but in the third phase of the research, subjects were re-tested with a modified version of the test material in order to compare their performance on the two different versions. The effect of history on test results is increased as the time between assessments increases (Ventry and Schiavetti, 1980:69), and therefore only a week was allowed between the first and second test.

❑ Testing or Test Practice Effects

If subjects had been subjected to previous testing with the same test or material, it is possible that their familiarity with the test could improve their scores (Struwig and Stead, 2001:137). This variable was particularly important to consider in the present study due to the redundancy of sentence material, which enables a person to recognise a sentence even if only one word is heard. If a subject is at all familiar with the sentence material, even fewer cues may be needed for correct recognition (Owens, 1983:359). For this reason, new subjects were used during each phase of the current project to ensure that no subjects had any prior exposure to the test material.

During Phase III, subjects were re-tested, but a week was allowed in between testing to reduce the practice effect. It should be noted, however, that the reason for re-testing the same subjects was to obtain an indication of the difference in performance under a “normal” test condition versus a condition where a hearing loss was simulated. However, the main goal of this procedure was to compare the inter-list variability in the normal and simulated loss condition for two different sets of lists (experimental and control), and not to establish the exact or absolute scores that could be expected under these two conditions. To ensure that the practice effect did not affect one group more than the other, the two groups were both re-tested after one week.

❑ Instrumentation

A common threat to internal validity in tests concerning hearing is the effect that changes in equipment or its calibration might have on results (Ventry and Schiavetti, 1980:73). To exclude this variable as an influence on results, data was collected over a limited period of time (approximately six months), and equipment was not moved during this period. All testing was conducted in the same sound-proof booth using the same equipment.

❑ Differential Selection of Subjects

Differences between subjects in the experimental and control groups may influence the validity of the research if these differences affect the performance of the subjects (Ventry and Schiavetti, 1980:77). Therefore, the selection criteria for both the experimental and control groups in Group 3 were identical and accounted for all the subject variables that might have an influence on performance. In addition, by using the same subjects for Experiment I and Experiment II in the third phase, a valid comparison could be made between the results of the two experiments, since the only difference between subjects in the two experiments was the simulated hearing loss introduced and controlled by the researcher.

❑ Mortality or Attrition

This term refers to the loss of subjects during the course of the research (Ventry and Schiavetti, 1980:78). This applies especially to follow-up studies where long-term effects of a disease or treatment program are monitored (Ventry and Schiavetti, 1980:79). In the present study, it was only the subjects in the third phase that had to be re-tested. The elapsed time between testing and re-testing was limited to one week in order to reduce the possibility of “losing” subjects. For ethical reasons, subjects were required to sign an informed consent form prior to participation,

which specified that they were allowed to withdraw from the study at any time. Fortunately, none of the subjects in the present study were lost in this way.

The second type of validity that was considered in planning the research is external validity. This type of validity relates to the extent to which the results can be generalised to one or more populations (Struwig and Stead, 2001:136). External validity depends in part on the internal validity of the study, but these two types of validity often pose conflicting demands to the researcher (Maxwell and Satake, 2006:157). This is due to the fact that greater control over threats to internal validity leads to less control over threats of external validity. If all the variables that could affect internal validity are controlled, it is likely that the results cannot be generalised to a larger population (Ventry and Schiavetti, 1980:82).

In some research studies, the extent to which data can be generalised is not considered a priority, as researchers are more focused on exploring a specific concept within particular boundaries (Ventry and Schiavetti, 1980:81-82). Within the present study, internal validity was controlled to a large extent, especially in terms of subject selection. All subjects fell within a specified age range and adhered to specific criteria in terms of auditory and language abilities, which limits the potential generalisation of results. However, since the research was aimed at the development of a new test, and was exploratory in this regard, the goal was not to obtain results that could be generalised to a larger population, but rather to refine the created test by carefully controlling as many variables as possible.

3.11 Conclusion

The current chapter described the research process followed to develop a valid and reliable Afrikaans test of sentence recognition thresholds in noise. The process consisted of three distinct phases, each yielding data that was

essential for the subsequent phase. The results of these procedures will be presented in the following chapter.

3.12 Summary

The main aim and sub-aims of the research were described in this chapter, followed by a discussion of the research design and research sample. Subsequently, the material and apparatus utilised in the execution of the different phases were described, followed by a revision of the ethical considerations of the research. A detailed description of the procedures of each of the three phases was provided and the chapter concluded with a discussion of the reliability and validity of the research project. The aims and procedures of all three phases are summarised in Table 3.13.

Table 3.13: Aims and procedures of each phase of the project

MAIN AIM: TO DEVELOP A VALID AND RELIABLE AFRIKAANS TEST OF SENTENCE RECOGNITION THRESHOLDS IN NOISE		
SUB-AIM	PROCEDURES	MOTIVATION
<p>Phase I</p> <p>Sub-aim: To develop a collection of recorded Afrikaans sentences suitable for the assessment of speech recognition in noise</p>	<ul style="list-style-type: none"> - Translate sentences from BKB (Bench and Bamford, 1979). - Compile additional sentences of similar length and structure using vocabulary from the Afrikaanse Reseptiewe Woordeskat test (Buitendag, 1994) and the Afrikaans Phonetically Balanced Word Lists for children 3-5 years. - Determine naturalness of sentences through submission to a panel of native Afrikaans speakers of different ages, occupations and educational backgrounds. - Describe grammatical and syntactical level of sentences. - Digitally record sentences. - Equate the average intensity level of all sentences and edit the recordings to eliminate unwanted silences. 	<ul style="list-style-type: none"> - BKB sentences adhere to requirements of suitable sentence material in terms of word familiarity, grammatical structure and memory effects (Olivier, 2000). - Additional material to be selected in order to enable the researcher to later eliminate sentences that are not suitable in terms of difficulty or validity, while still retaining a large enough sample to compile a large number of lists (Vaillancourt et al., 2005). - Naturalness and readability to be assessed in order to determine cultural and linguistic validity as well as degree of difficulty in terms of vocabulary (Vaillancourt et al., 2005). - Sentences to be pre-recorded in order to yield a standardised presentation method (Rupp, 1980).
<p>Phase II</p> <p>Sub-aim: To select, from the recorded material, a collection of sentences with equivalent intelligibility in the presence of noise</p>	<ul style="list-style-type: none"> - Generate a background noise spectrally matched to the recorded speech sample. - Present sentences in the presence of the generated noise at an SNR of -5 dB to a group of normal-hearing native speakers of Afrikaans (adults). - Determine the mean percentage of syllables repeated correctly in each sentence across listeners. - Select sentences with similar (mean +/- standard deviation) percentage scores and eliminate outlying sentences from the sample. - Present each of the remaining sentences at two different SNRs (-2 dB and -8 dB) to a new group of normal-hearing Afrikaans speaking adults. - Determine the percentage of syllables correct in each sentence at each of the different SNRs. - Determine the intelligibility slope of each of these sentences. - Identify and select the sentences with similar intelligibility slopes. 	<ul style="list-style-type: none"> - Phonetic content, word familiarity as well as variations in intonation influence intelligibility in noise, so presenting sentences in a spectrally matched noise provides a means of selecting sentences of similar intelligibility (Nilsson et al., 1994). - Previous studies with similar aims have found a 50% recognition score at an SNR of approximately -5 dB. - Presenting sentences at different SNRs shows the sensitivity of each sentence to an alteration in the test condition (i.e. level of noise), thereby ensuring that all sentences provide a sensitive measure of speech recognition in noise.

Table 3.13: Aims and procedures of each phase of the project (continued)

<p>Phase III</p> <p>Sub-aim:</p> <p>To compare inter-list reliability and response variability of two list sets compiled using two different methods of list compilation</p>	<ul style="list-style-type: none"> - Describe the phonetic content of the total collection of sentences and arrange them in phonetically balanced lists of ten sentences each. - Create lists of ten sentences each with similar predicted intelligibility slope and SNR-50 according to performance in previous phases. - Determine the mean SNR-50 (and standard deviation) of each set of lists in a group of normal-hearing Afrikaans speaking adults using an adaptive presentation method. - Eliminate from the collection any lists that yield mean thresholds and/or response variability that differ significantly from the other lists. - Compare the inter-list reliability and response variability attained through both methods of list compilation. - Retest the same subjects, but with a simulated high frequency hearing loss, with the same lists and compare the inter-list equivalence of the two sets of lists in the filtered (simulated loss) condition. 	<ul style="list-style-type: none"> - Two methods of arranging sentences into lists of ten followed in order to determine whether there is a more time-efficient yet reliable manner than the traditional “phonetic balance method” to obtain equivalent lists.
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4. RESULTS

4.1 Introduction

The development of an Afrikaans test for sentence recognition in noise required several data collection and analysis procedures. These procedures were conducted in three separate phases, as described in the previous chapter. The results of each of the three phases of data collection are presented separately in this chapter.

4.2 Phase I

The aim of the first phase was to develop an appropriate collection of recorded Afrikaans sentences for the assessment of speech recognition in noise. This section will provide an exposition of the results attained during this phase. It will provide a description of the sentence material compiled, along with the results of the naturalness and grammar ratings.

4.2.1 Compilation of sentence material

The collection of sentences was compiled from three different sources, as described in the methodology. These three sources are the BKB sentences (Bench and Bamford, 1979), the “Afrikaanse Reseptiewe Woordeskattoets” or ARW (Buitendag, 1994) and the “Foneties gebalanseerde Woordelyste” for children 3-5 years or PBC lists (Phonetically Balanced Children’s word lists) (Tesner and Laubscher, unpublished). Figure 4.1 below indicates the composition of the collection, showing the percentage of sentences derived from each of the three sources.

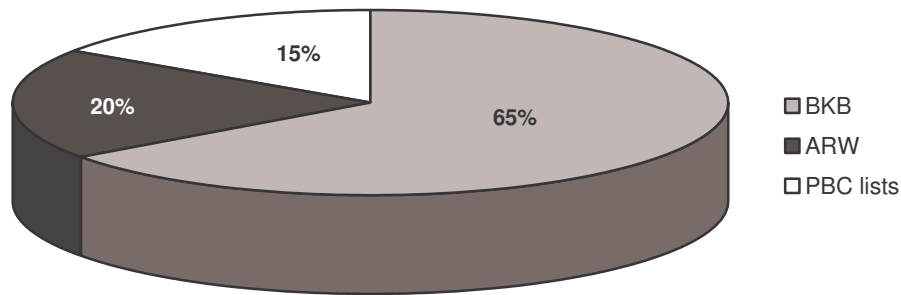


Figure 4.1: Percentage of sentences derived from each source.

As shown in the figure, the majority of sentences (65%; $n = 336$) were translations of the BKB sentences. 20% of the total collection ($n = 102$) was made up by sentences compiled from the ARW, and 15% ($n = 80$) from the children’s word lists. The total number of sentences was 518. These sentences had a mean length of 5.6 words or 7.1 syllables. The length of the sentences derived from the three different sources is shown in Table 4.1 below.

Table 4.1: Length of sentences derived from each source

NUMBER OF WORDS				
	MINIMUM	MAXIMUM	MEAN	STD DEV
BKB	4	8	5.4	0.93
ARW	4	8	6	0.93
PBC lists	5	8	5.9	0.84
NUMBER OF SYLLABLES				
	MINIMUM	MAXIMUM	MEAN	STD DEV
BKB	4	9	7	1.17
ARW	5	9	7.5	1.1
PBC lists	5	9	6.7	1.11

As shown in the table, the minimum and maximum number of words and syllables remained fairly constant across all three sources, with slight variability in terms of the mean number of words and syllables. The goal of

limiting this variability was to facilitate uniformity within the total collection of sentences. The uniformity in length as well as the naturalness of the translated BKB sentences were facilitated by changing the content of some of these sentences, as described under the methodology of this phase. A total number of 113 sentences were altered, as shown in Table 4.2 below.

Table 4.2 Changes made to BKB sentences in translation

TYPE OF CHANGE	MOTIVATION	NO. OF CHANGES
Semantical	Direct translation would result in sentences that are too long or would sound artificial/unnatural	49
Tense / time	Direct translation retaining the same tense would make sentences too long	34
Cultural	Direct translation would contain vocabulary or concepts that are uncommon to South Africans	30

As shown in the table, three types of changes were made. The majority of changes were semantical, and were mainly made to control the length and improve the naturalness of the sentence. Changes in tense or time were mostly necessary to limit the length of the sentence, and cultural changes were made to eliminate concepts that would be foreign to South African listeners. These changes related specifically to concepts such as food, weather, transport systems and other aspects of everyday living (such as the postman or milkman) that are foreign to the South African context.

4.2.2 Rating of naturalness

The total collection of 518 sentences that were submitted to the first half of Group A of the participants ($n = 5$) received a mean rating of 6.9 with a standard deviation of 0.2. Only 3 sentences in the collection scored a mean rating equal to or lower than 6.0. Due to the fact that so few sentences received a mean rating below 6, sentences that received a rating below six from at least two candidates were also reviewed. Although the mean rating for these sentences was still above the initially determined criterion of 6, it was decided to adapt these sentences according to the recommendations of the

participants, as two out of the five participants felt that these sentences should be changed.

Two additional sentences were modified. The first sentence was changed slightly because it was found to be too long only after sentences had been submitted for rating. The second sentence (a translation of “The boy had a toy dragon”) was objected to by one of the subjects not on the part of naturalness, but rather for its poor predictability in comparison to the rest of the material. This objection was considered and the content of the sentence simplified (to “The boy had a red car”). The different reasons for reviewing sentences and the number of sentences reviewed for each reason are summarised in Table 4.3.

Table 4.3: Reasons for revision of sentences

REASON FOR REVISION	NO. OF SENTENCES
Mean rating \leq 6.0	3
Rating < 6 from two or more subjects	8
Sentence too long	1
Content unpredictable	1
Total number of sentences revised	13

These 13 sentences that required revision after the first round of rating were corrected as follows.

1. If the recommendations for change made by the subjects were identical, the sentence was changed to their suggestion. Four sentences were changed in this manner.
2. If two or more subjects recommended a change that differed slightly between subjects, the suggestions were combined to alter the sentence. Six sentences were edited in this way.
3. If the subjects’ suggestions were incompatible, the sentence was rejected from the collection. Only one sentence had to be rejected for this reason.

4. One sentence was changed because it was found to be too long, and one sentence was changed because one of the subjects considered its content to be unpredictable.

In total, twelve sentences were altered and one sentence rejected in the first round. The twelve modified sentences were then submitted to an additional five native Afrikaans speakers (subjects 6 – 10 of Group A, as shown in Table 3.4). The mean rating for these twelve sentences in the second round was 6.5, and only 2 sentences received a mean rating lower than six (5.4 and 5.6). These two sentences were therefore excluded from the collection. None of the other sentences received a rating lower than six from more than one participant. After two rounds of naturalness rating, therefore, 515 sentences remained. These sentences were subsequently submitted to an expert in language development and analysis for a rating of their grammatical level.

4.2.3 Rating of grammatical level

The complexity of the grammatical structure of each sentence was rated on a 7-point scale, as described under 3.7.3. The total number of sentences allocated each of the 7 possible ratings is illustrated in Figure 4.2.

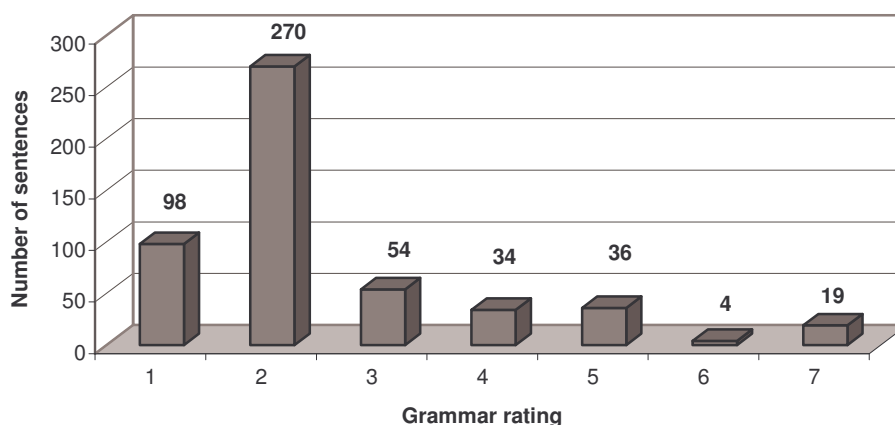


Figure 4.2: Number of sentences allocated each of the different grammar ratings (1 = simplest; 7 = most complex)

As shown in Figure 4.2, the majority of sentences (81%) received a rating between 1 and 3. Only 93 sentences (19% of the total collection) received a grammar rating between 4 and 7. These distributions are also indicated in Figure 4.3, which clearly shows that the majority of sentences received one of the first three ratings (1, 2, or 3, allocated to 19%, 52% and 10% of the collection respectively).

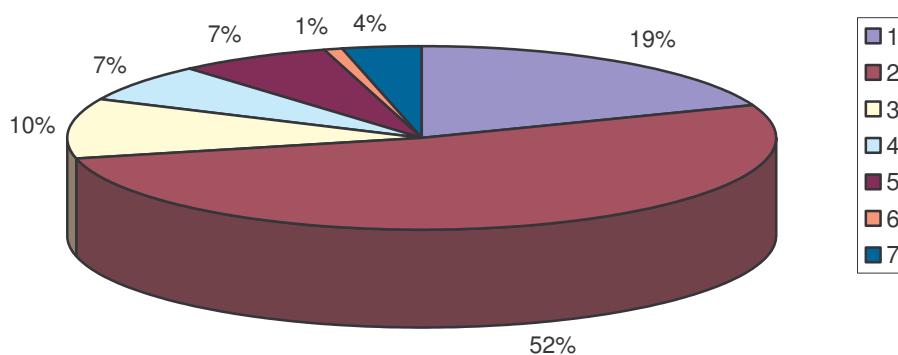


Figure 4.3: Distribution of grammar ratings across 515 sentences

Since all these ratings fell within the 3-5 year age range, it was decided not to reject any sentences on the basis of their grammar rating. These ratings nevertheless provided valuable basic data to describe the difficulty of the sentence collection, which was also used in subsequent phases for analyses of the sentence material.

4.3 Phase II

The second phase was aimed at selecting, from the material recorded in Phase I, a collection of sentences with equivalent intelligibility in the presence of noise. Two consecutive equalisation procedures were followed to attain this

aim, and the results of these two procedures will be presented separately in this section.

4.3.1 First equalisation procedure: Selection of equivalent subset of sentences

The first procedure yielded the percentage of syllables discerned correctly for each sentence by each participant at an SNR of -5 dB. Using these results, a mean percentage could be calculated for each sentence, as well as the overall mean of all the sentences. The mean percentage of syllables discerned correctly by all participants for all sentences at SNR-5 was 52%, very close to the predicted 50% for this SNR. The standard deviation from this mean was 27%. Sentences were selected from the collection if their mean score across participants fell within one standard deviation from the mean, i.e. between 25% and 78%, as shown in Figure 4.4 below.

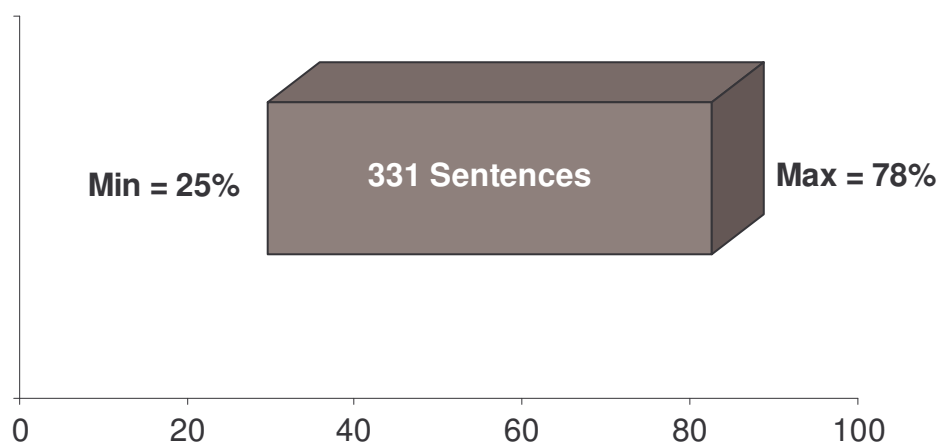


Figure 4.4: Sentences selected after first equalisation procedure according to intelligibility scores at SNR-5 dB

As shown in the graph, a total of 331 sentences fell within these limits. One sentence was excluded as its content was found to be identical to another in

the collection. Therefore, 330 sentences were selected for use in the second equalisation procedure.

These findings provided essential data required for the commencement of the second procedure. However, additional findings were also made that were not essential for subsequent procedures, but could nevertheless guide the methods followed in these procedures. The first finding of this kind was the gender differences in obtained scores. The mean score for male participants across all sentences was 47%, whereas female participants scored a mean of 56%. There was also a marked difference between the minimum and maximum scores of the two genders. This difference is indicated in Figure 4.5 below, where the values at the beginning and the end of the bar indicate the minimum and maximum scores respectively.

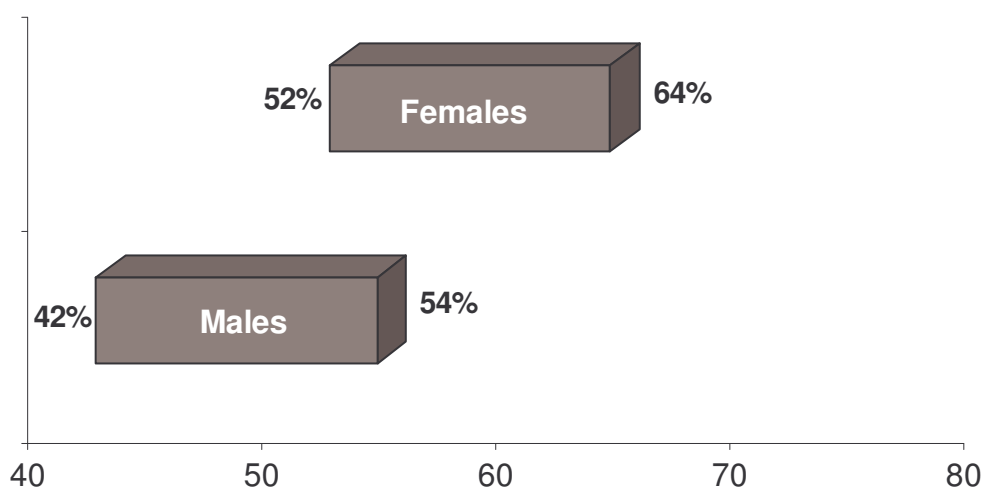


Figure 4.5: Gender differences as demonstrated by results from 1st equalisation procedure

As demonstrated in Figure 4.5, there was a clear difference in performance between the two genders, with female participants performing significantly better than male participants. This finding was confirmed with the statistical analysis of variance, which demonstrated that the gender difference was

significant at a level of $p < 0.0001$ (F value = 15.47). The immediate implication of this gender difference was that gender had to be controlled in the second equalisation procedure by ensuring that an equal number of male and female participants were used in this procedure.

The grammar rating awarded to each sentence during Phase I was also compared to the mean percentage score of the sentence in the SNR-5 condition (first equalisation procedure). Sentences were grouped according to their grammar rating (seven groups, namely 1 to 7), and the mean intelligibility score of each group was calculated. From this comparison it was clear that the percentage score did not reflect the grammatical complexity of the sentence, as demonstrated in Figure 4.6, since sentences with greater grammatical complexity often obtained scores much higher than less complex sentences.

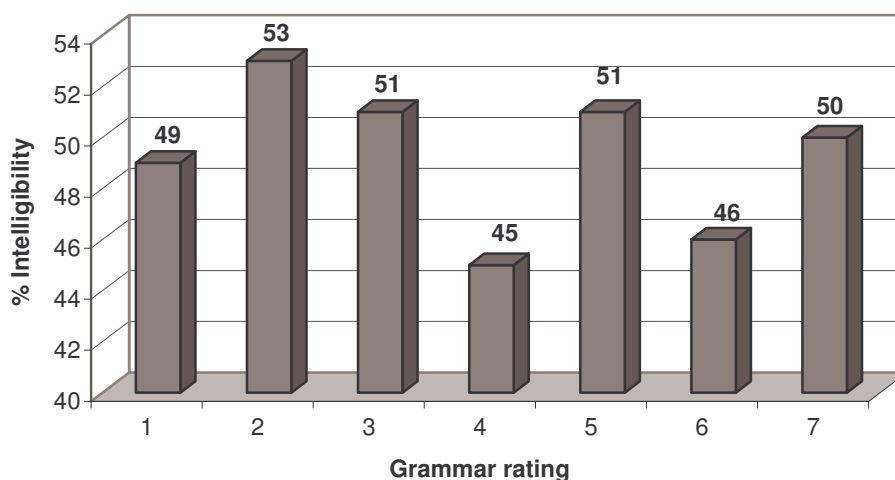


Figure 4.6: Mean percentage intelligibility score as a function of grammar rating

As shown in Figure 4.6, grammar ratings did not appear to have an effect on intelligibility scores, since complex sentences with a grammar rating of 7 scored a much higher average in terms of intelligibility than less complex sentences (for example sentences with a grammar rating of 4). The statistical

analysis of the data also revealed that sentences with a similar grammar rating sometimes had a significant difference in intelligibility scores, whereas sentences with large differences in terms of grammar rating sometimes had very similar intelligibility scores. Table 4.4 below shows that there were three instances where adjacent grammar ratings (for example 1 and 2) differed significantly in terms of intelligibility.

Table 4.4: Differences in intelligibility scores between grammar ratings

GRAMMAR RATINGS	ANOVA RESULT	SIGNIFICANT (< 0.05)?
1 and 2	0.003	Yes
2 and 3	0.209	No
3 and 4	0.016	Yes
4 and 5	0.021	Yes
5 and 6	0.415	No
6 and 7	0.549	No

Another finding that was explored was the apparent practice effect that was observed across participants. To counterbalance the possible effect of this aspect during testing, each participant started with a different list. However, with the data rearranged according to the order in which the lists were presented to the participants, there appeared to be a definite improvement in performance within the first ten to twenty sentences. The mean performance of the subjects on the first and all subsequent sentences presented to them are shown in Figure 4.7.

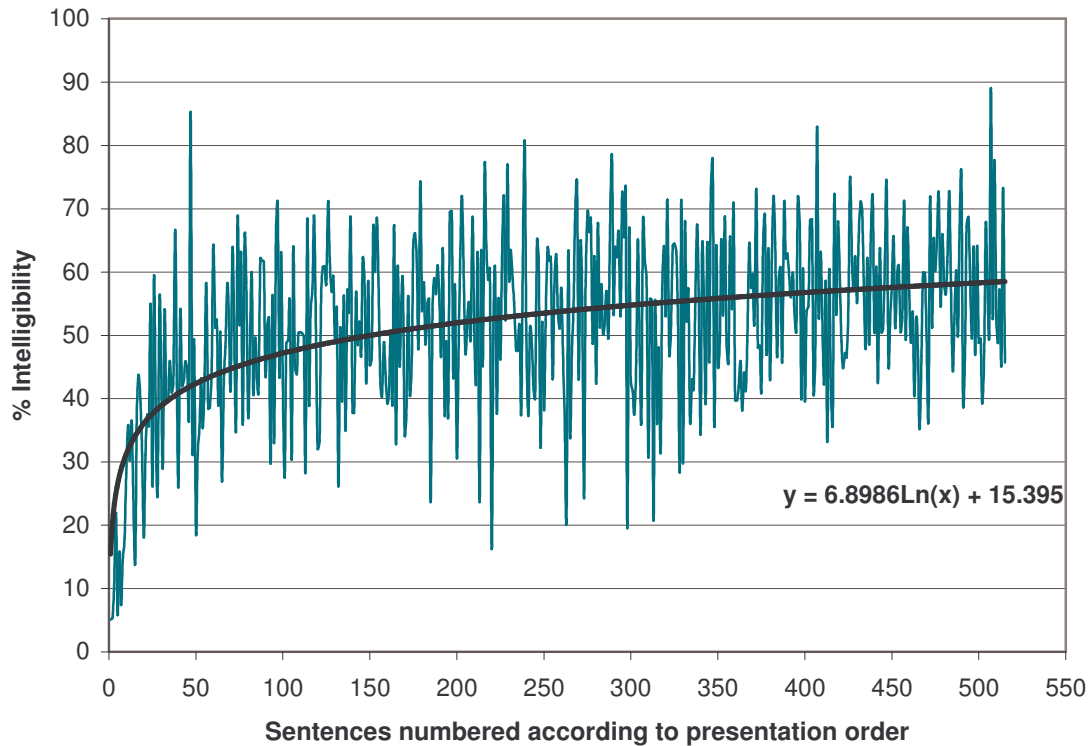


Figure 4.7: Mean performance of subjects arranged according to presentation order, showing apparent trend of improvement

The logarithmic trendline (calculated as $y = 6.8986\ln(x) + 15.395$) appears to indicate an improvement in percentage intelligibility within the first 50 sentences. However, statistical analysis showed that there were no significant differences in performance within the first twenty sentences. A Friedman two-way analysis of variance (ANOVA) test resulted in a p-value > 0.05 for the first ten, second ten and first twenty sentences, which indicates that these differences are insignificant, as shown in Table 4.5.

Table 4.5: ANOVA results of practice effect for SNR-5 test condition

SENTENCES COMPARED	p-VALUE	SIGNIFICANT (< 0.05)?
1 st – 10 th	0.94	No
11 th – 20 th	0.84	No
1 st – 20 th	0.24	No

Despite the fact that these differences were not statistically significant, the possibility of a practice effect was still controlled for in subsequent test procedures by presenting a practice list to subjects prior to testing.

4.3.2 Second equalisation procedure: Selecting sentences with similar intelligibility slopes

During the second equalisation procedure, the 330 sentences selected as a result of the first, were presented to 12 subjects (6 male, 6 female) at a fixed SNR. The first 6 subjects listened to the sentences at an SNR of -8 dB, whereas the last 6 subjects heard the sentences at an SNR of -2 dB. Percentage intelligibility score per sentence was again calculated as the percentage of syllables correctly repeated. The mean percentage for all sentences across all subjects at SNR-8 dB was found to be 18.2%, with a standard deviation of 12.7%. At an SNR of -2 dB, the mean percentage was found to be 89.5%, with a standard deviation of 11.0%. These values were used in conjunction with the scores obtained at SNR-5 to determine the intelligibility slope of each of the 330 sentences, and were subsequently used to select a number of sentences with similar intelligibility slopes, as described in section 3.8.2. The mean score at each SNR, along with the differences between the different SNR scores are depicted in Figure 4.8.

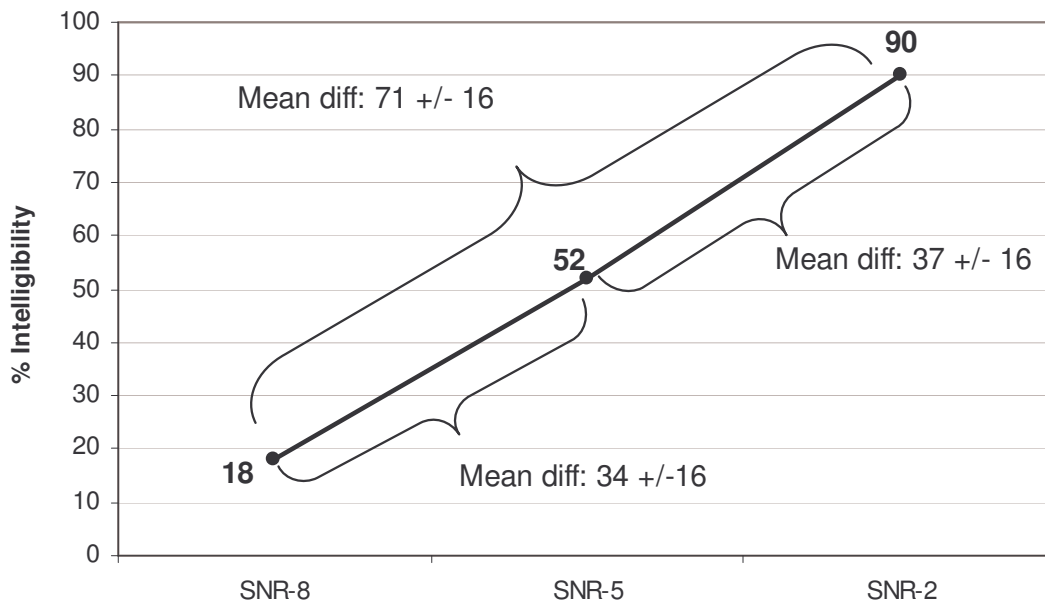


Figure 4.8: Mean scores at each SNR. Mean differences and standard deviation of these differences (indicated by \pm) are also shown.

As shown in Figure 4.8, the mean difference between the scores at SNR-2 and SNR-8 was 71%, with a standard deviation of 16. Therefore, sentences with a mean difference between the SNR-8 and SNR-2 scores that fell between 55 and 87% were selected for the collection, provided that the difference between their scores at SNR-2 and SNR-5 was between 5 and 68% (i.e. within 2 standard deviations of the mean) and the difference between their SNR-5 and SNR-8 scores fell between 3 and 66% (also within 2 standard deviations of the mean difference between these points). These procedures led to the selection of 222 sentences that fell within the stipulated criteria, for use in the final phase of the project.

The effect of gender on performance was again investigated. The mean score of female participants was again higher under both conditions (SNR-8 and SNR-2). The differences between male and female averages for both conditions are illustrated in Figure 4.9.

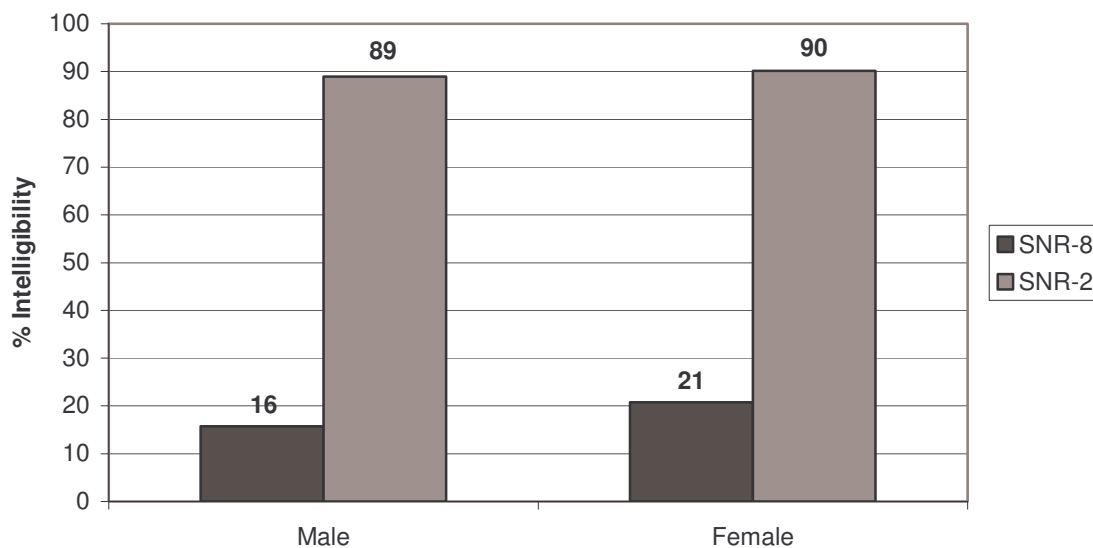


Figure 4.9: Mean scores for male and female participants at SNR-8 and SNR-2

Although there was a slight difference between male and female scores for both conditions, this difference was found to be insignificant (with $p = 0.1$ at SNR-8 and $p = 0.23$ at SNR-2). Despite the absence of the gender effect here, it was still decided to control this variable in the final phase (by using an equal number of male and female participants) in light of the findings of the first procedure.

The correlation between the grammar rating and intelligibility scores attained in the second procedure was also explored. These results are indicated in Figure 4.10. As indicated in the graph, there was no clear correlation between the grammatical complexity and the intelligibility score for either SNR condition. Statistical analyses of the findings confirmed this observation, since the correlation was found to be insignificant (0.11 for SNR-8 and 0.13 for SNR-2).

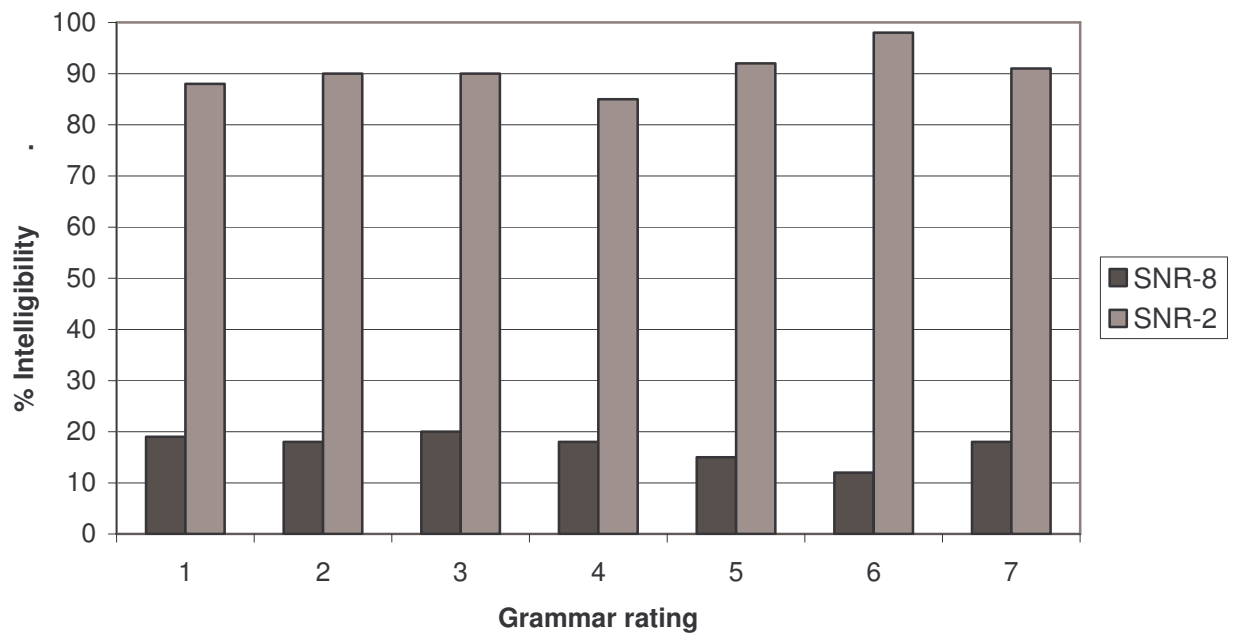


Figure 4.10: Intelligibility scores for SNR-2 and SNR-8 as a function of grammar rating

The effect of practice on performance was also examined in the second procedure. Scores obtained by each participant were rearranged according to the order in which the playlists were presented to them. Figure 4.11 below indicates the mean scores of subjects arranged according to the order in which sentences were presented to them.

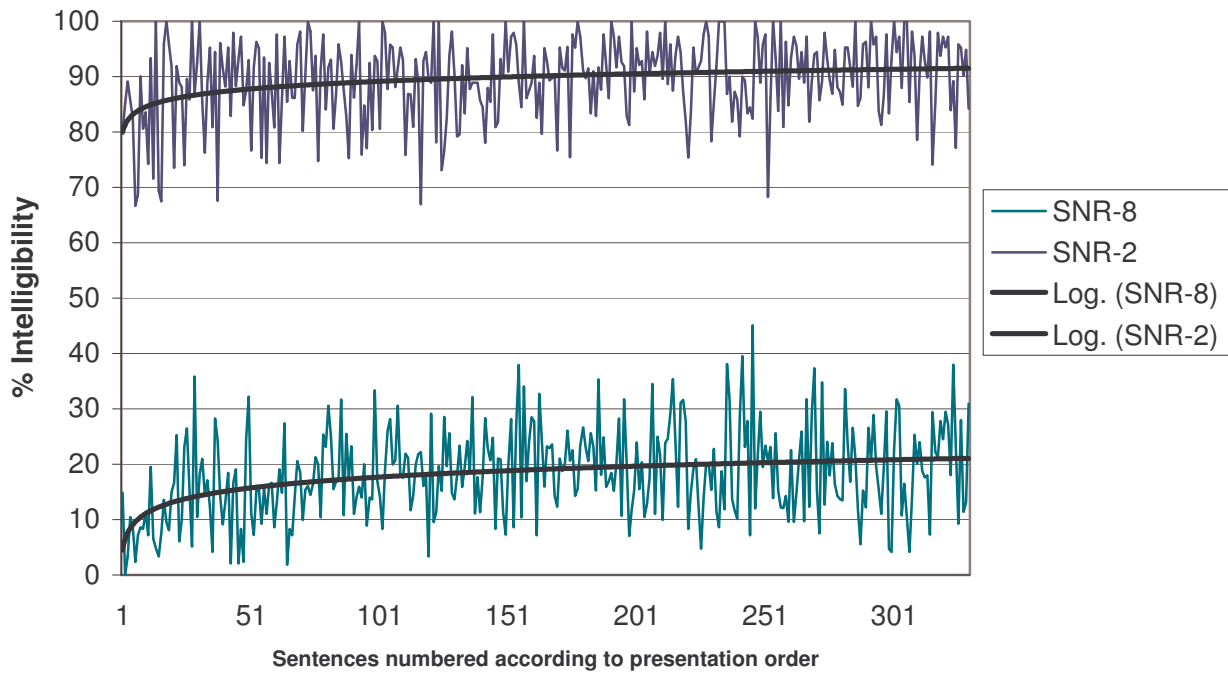


Figure 4.11: Mean performance at SNR-2 and SNR-8 arranged according to presentation order. Trendlines indicate practice effect.

As shown in this graph, the apparent practice effect (indicated by the trendline sloping upwards slightly) seems to be less than that noticed during the first equalisation procedure (see Figure 4.7). To confirm this observation, a Friedman two-way analysis of variance was also conducted on these findings. The results of the analysis are indicated in Table 4.6.

Table 4.6: ANOVA results of practice effect for SNR-8 and SNR-2 test conditions

SENTENCES COMPARED	p-VALUE	SIGNIFICANT DIFFERENCE ($p < 0.05$)?
SNR-8 1 st to 10 th	0.94	No
SNR-8 11 th to 20 th	0.97	No
SNR-8 1 st to 20 th	0.99	No
SNR-2 1 st to 10 th	0.84	No
SNR-2 11 th to 20 th	0.14	No
SNR-2 1 st to 20 th	0.26	No

As shown in the table, the overall comparison of the 1st to 10th, 11th to 20th and 1st to 20th sentences presented under both the SNR-8 and SNR-2 conditions

revealed no significant differences between subjects' scores. Multiple comparisons were also conducted to determine if there was a significant difference between any two sentences in terms of subject performance. The p-value for these differences was also above 0.05 for each pair compared, and no significant differences were therefore found.

4.4 Phase III

The aim of the third phase was to compare inter-list reliability and response variability for two different methods of list compilation. Three main processes were followed to achieve this aim, namely list compilation and two different experimental applications of the lists. The results of each of these three processes are provided in this section.

4.4.1 List compilation

Two methods of list compilation were followed, as described in Chapter 3 (section 3.9.1). The results of these two methods will first be discussed separately, and will then be compared.

4.4.1.1 Slope lists (experimental method)

The first set of lists ("slope lists") consisted of 22 lists with 10 sentences each. The first 220 sentences of the total of 222 sentences carried over from Phase II were used to compile these lists. These sentences were arranged into lists according to the mean intelligibility slope of each list (defined as the mean slope of all the sentences in the list) by exchanging sentences between lists as described in section 3.9.1.1. The resulting slopes after exchanges are shown in Figure 4.12.

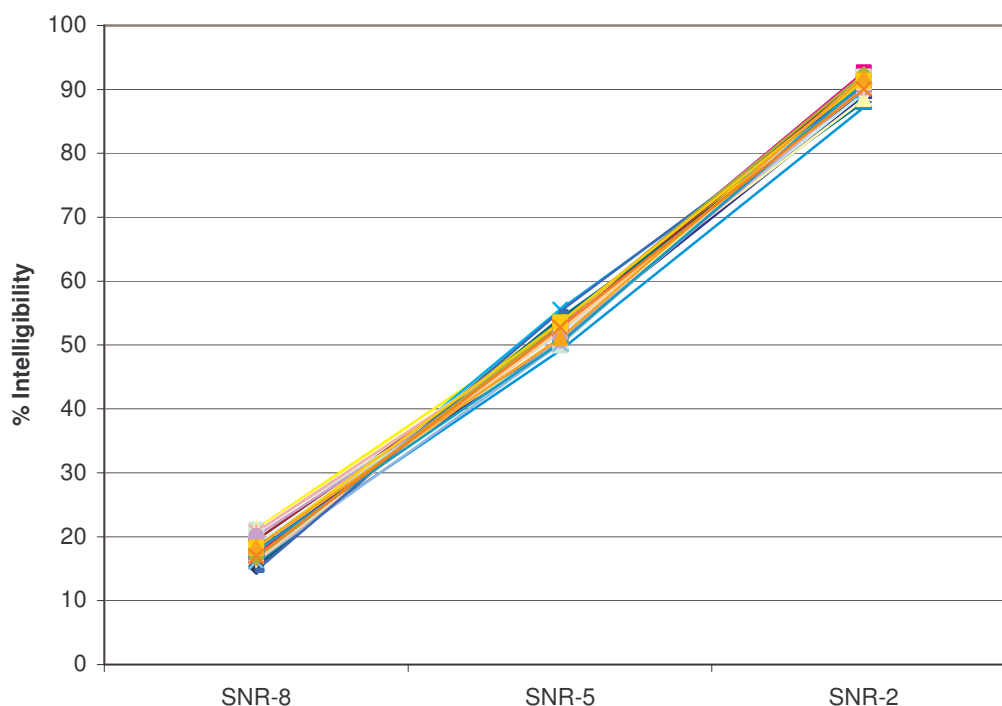


Figure 4.12: Intelligibility slope of each of the slope lists after exchange of 14 sentences

After the exchange of 14 sentences between the lists (that were at first grouped merely based on numerical order), the range between the minimum (lowest scoring list) and maximum (highest scoring list) at each point on the slope was down to 6 or 7 dB, with a standard deviation of 2 or lower. These findings are summarised in Table 4.7.

Table 4.7: Means and range of intelligibility scores of slope lists at each SNR

	SNR-8 SCORE (%)	SNR-5 SCORE (%)	SNR-2 SCORE (%)
Mean	18	52	90
Standard deviation	2	2	1
Minimum	15	49	87
Maximum	21	56	93
Range	7	6	6

Since it was only necessary to exchange 14 sentences in order to yield such a small variability, it was possible to compile a highly equivalent set of lists (based on previous experimentation) within a short amount of time.

4.4.1.2 Phonetically balanced lists (control method)

The control method of list compilation resulted in 22 phonetically balanced lists consisting of 10 sentences each. Within the total sentence collection 49 different phonemes occurred (see Table 3.12 for frequency of occurrence of each phoneme). For each of these 49 phonemes, an ideal count per list was determined (as the total number of occurrences, divided by 222 sentences and multiplied by 10). The phonetic balancing procedure was aimed at arranging sentences into lists in such a way that the occurrences of all the phonemes in all the lists were as close as possible to their ideal occurrence. Because it was not possible to attain an exact balance, the goal was to attain an optimum arrangement of the lists so that each list had the smallest amount of errors and the lowest possible total error value, without jeopardising the balance of another list.

The total number of errors for each list was calculated as the number of times that a phoneme deviated one or more from the ideal count, and included both positive and negative errors. These values are depicted in Figure 4.13. The minimum number of errors per list was 8 (Lists 13 and 22) and the maximum number of errors was 18 (List 9). The average number of errors per list was 11.8.

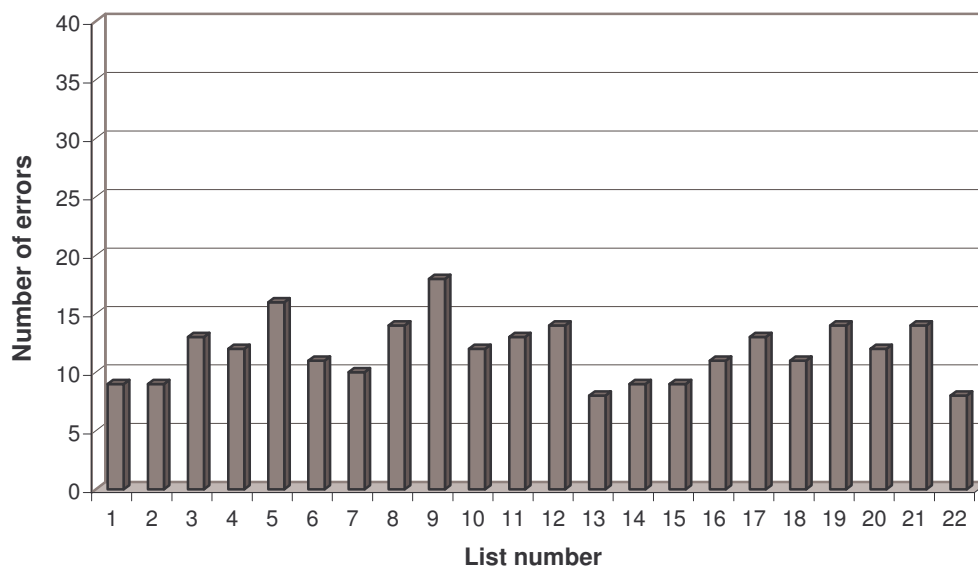


Figure 4.13: Number of errors in phonetic balance per PB list

The total number of phoneme counts for all the lists was 1078 (22 lists x 49 phonemes). The errors on each of these counts were calculated as the difference between the ideal occurrence and the actual occurrence. The resulting values were then rounded to the closest 0.5, and it was thereby possible to determine the percentage of counts that deviated from exact balance with a particular number of phonemes. These percentages are depicted in Figure 4.14.

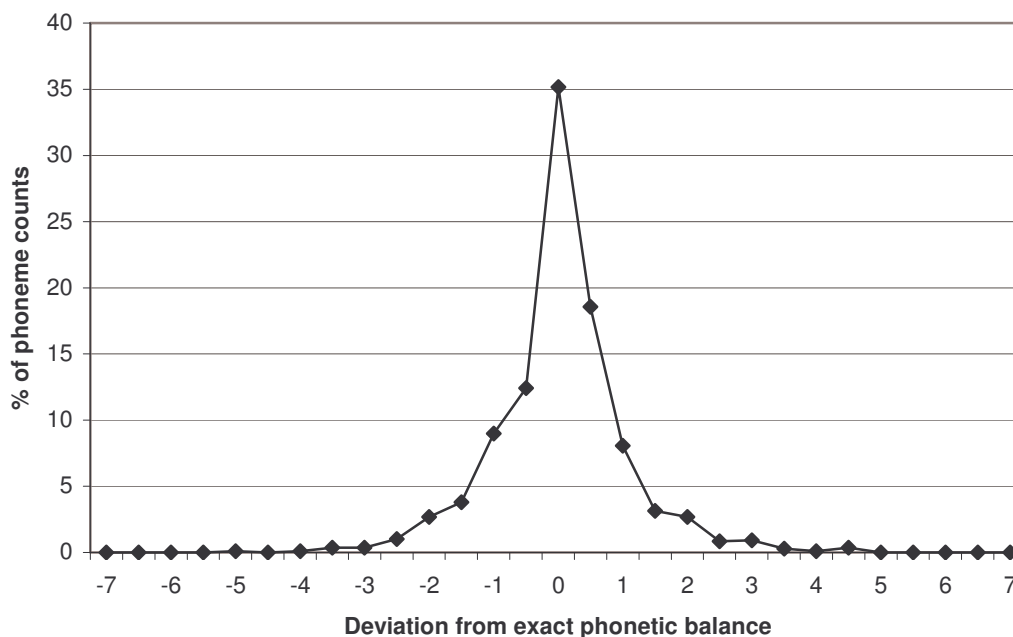


Figure 4.14: Deviation of phoneme counts for all PB lists as a function of the total phoneme count (all phonemes across all sentences; n = 1078)

As illustrated in the graph, 379 (35%) of these counts showed an error value of 0, that is, these phoneme counts were exactly equal to their ideal occurrence. Of the total number of errors for all phonemes across all lists, 83.2% fell within +/- 1 phoneme of its ideal occurrence. The maximum error for a single phoneme was + 5 (shortage of 5), but this occurred only once (less than 0.1% of total phoneme counts).

4.4.1.3 Comparison of two methods

As described under section 3.9.1 in the previous chapter, the slope lists were compiled purely according to the mean intelligibility slopes of each list, without any consideration to phonetic content. The composition of the PB lists, in turn, was based solely on the phonetic content, and without consideration to intelligibility slopes. However, after these lists had been compiled, it was possible to determine the intelligibility slopes of the PB lists, as well as the phonetic content of the slope lists and compare the lists accordingly.

In terms of phonetic balance, the slope lists exhibited a much larger number of errors per list, as shown in Figure 4.15 below. In this set of lists, the minimum number of errors per list was 21 (list 13) and the maximum was 37 (List 8). The mean number of errors per list was 28.8, which was much higher than the mean number of errors for the PB lists (11.8).

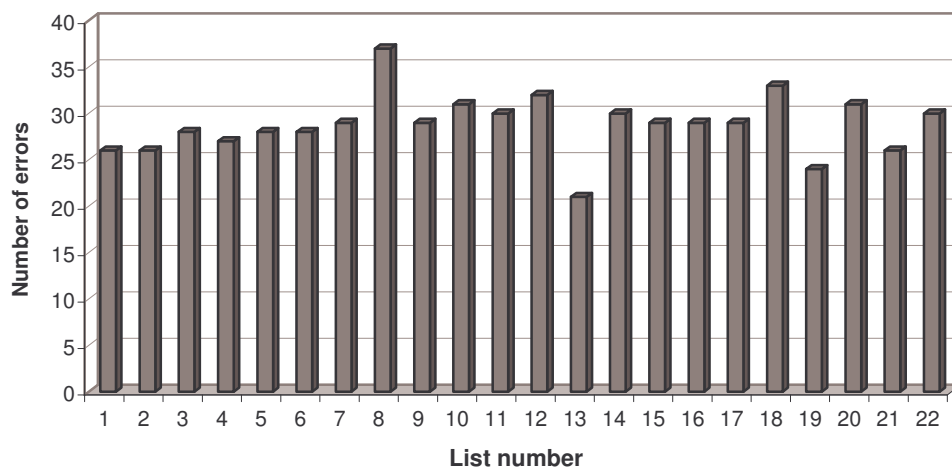


Figure 4.15: Number of errors in phonetic balance per slope list

The total number of phoneme counts for all the lists was 1078 (22 lists x 49 phonemes). As with the PB lists, the phonetic errors on each of the 1078 phoneme counts were calculated as the difference between the ideal occurrence and the actual occurrence. The resulting values were then rounded to the closest 0.5, and the percentage of counts that deviated from exact balance with a particular number of phonemes could be determined. These percentages are depicted in Figure 4.16.

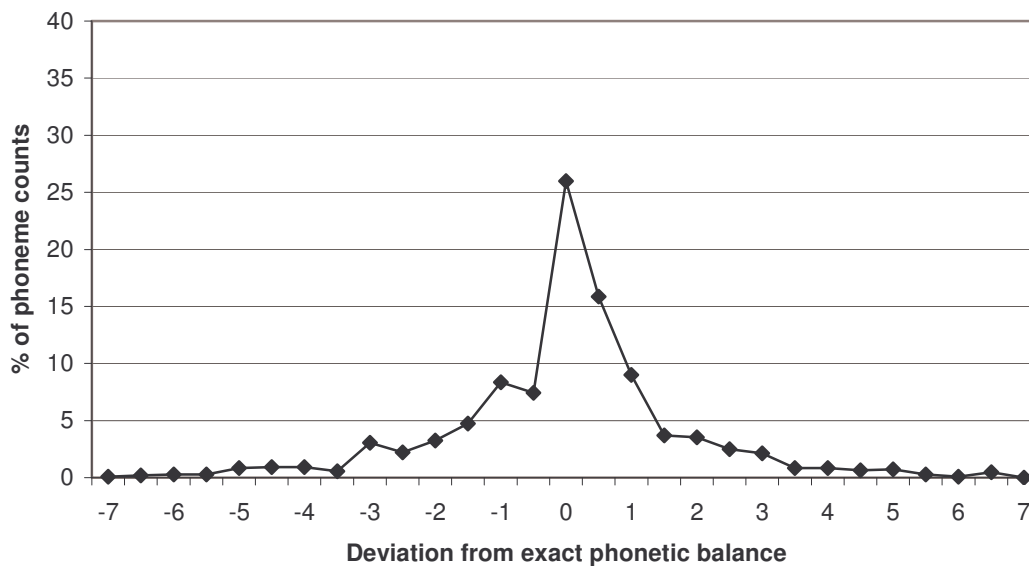


Figure 4.16: Deviation of phoneme counts for all slope lists as a function of the total phoneme count (all phonemes across all sentences; n = 1078)

As demonstrated in Figure 4.16, only 26% of the phoneme counts showed a deviation of 0 phonemes. A total of 66.6% of the counts differed with only +/- one phoneme from its ideal occurrence or exact phonetic balance. The maximum error on a single phoneme was -12 (an excess of 12 phonemes), which occurred once. Errors equal to or larger than +/-5 occurred a total of 38 times (3.5% of the total phoneme counts).

An additional comparison between the slope lists and PB lists was made by calculating the intelligibility slopes of the PB lists. These slopes are shown in Figure 4.17 below.

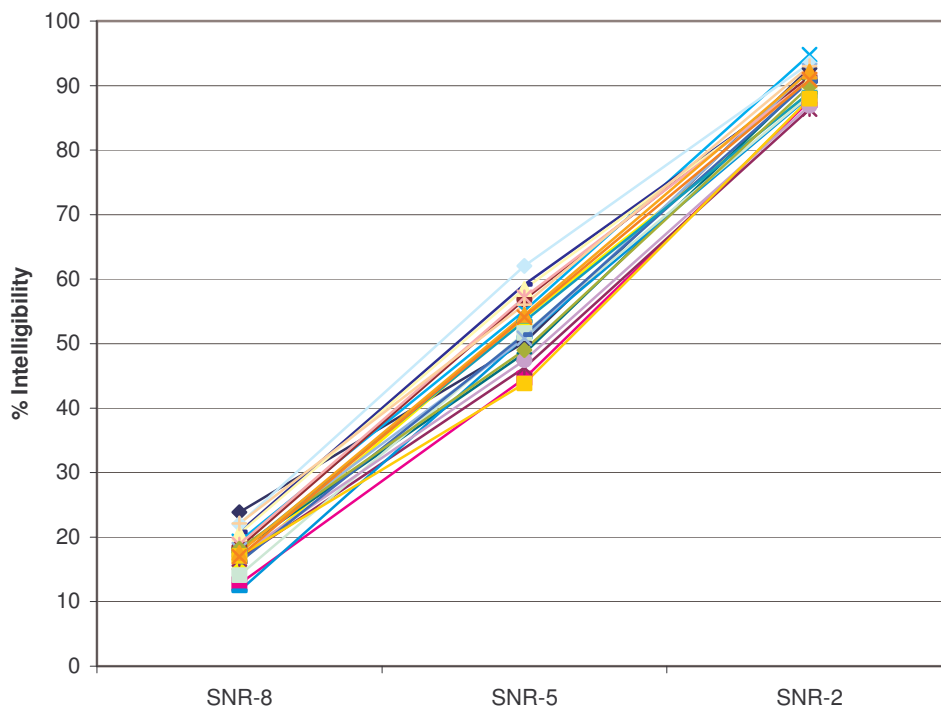


Figure 4.17: Intelligibility slopes of PB lists

The variability within this collection can be seen by the differences between the minimum and maximum scores at each point on the slope. These values are indicated in Table 4.8, along with the mean and standard deviation at each SNR.

Table 4.8: Means and range of intelligibility scores of PB lists at each SNR

	SNR-8 SCORE (%)	SNR-5 SCORE (%)	SNR-2 SCORE (%)
Mean	18	53	90
Standard deviation	3	5	2
Minimum	12	44	86
Maximum	24	62	95
Range	12	18	8

The range between the minimum and maximum was greater than for the slope lists, and similar to the ranges found when the slope lists were arranged in numerical order prior to the exchanges that were necessary to attain

equivalent slopes (see Figure 3.5), namely 10%, 17% and 11% for SNR-8, SNR-5 and SNR-2 respectively.

4.4.2 Experiment I: List application in group of normal-hearing listeners

After the compilation of the two sets of lists, these lists were subjected to an experimental application. The first experiment used a group of normal-hearing young adults with a negative otologic history as subjects. The experimental (slope) and control (PB) lists were both used in this experiment, and subjects were randomly assigned to either the experimental or control group. The results of the experiment were then used to determine the mean score (SNR-50 threshold) for each participant across lists, as well as the mean score for each list across participants. Results obtained from the application of the slope lists and PB lists are presented separately in this section.

4.4.2.1 Slope lists

The mean (average) SNR-50 of the 22 slope lists across subjects ($n=10$) was -2.6 dB, with a standard deviation of 0.45 across subjects. The mean SNR-50 thresholds and standard deviation thereof for each subject across all lists are depicted in Table 4.9.

Table 4.9 Means and standard deviations of SNR-50 for each subject across the 22 slope lists (unfiltered)⁸.

SUBJECT NO.	MEAN	STD DEV
1	-2.52	1.05
2	-3.27	0.98
3	-2.94	1.09
4	-2.45	1.16
5	-2.94	1.45
6	-2.15	0.90
7	-2.21	1.02
8	-2.73	1.14
9	-1.85	1.43
10	-2.97	1.23
Maximum	-1.85	1.45
Minimum	-3.27	0.90
Overall mean	-2.60	1.14

As indicated in Table 4.9, the maximum value (poorest score, as indicated by the better SNR level required for 50% performance) was -1.85 dB. The minimum value (best performance) was -3.27 , resulting in a performance range of 1.42 dB across subjects. The within-subject variability is indicated by the standard deviation in performance for each subject. As the standard deviations ranged from 0.9 to 1.45 dB, the largest deviation within a subject was still below 1.5 dB.

The mean score for each of the 22 slope lists across subjects was also calculated. These scores, along with standard deviations, are depicted in Figure 4.18.

⁸ The results for the first experiment with normal-hearing listeners refers to the lists as “unfiltered” to avoid confusion with the results for the second experiment, where lists were filtered to simulate a high frequency hearing loss.

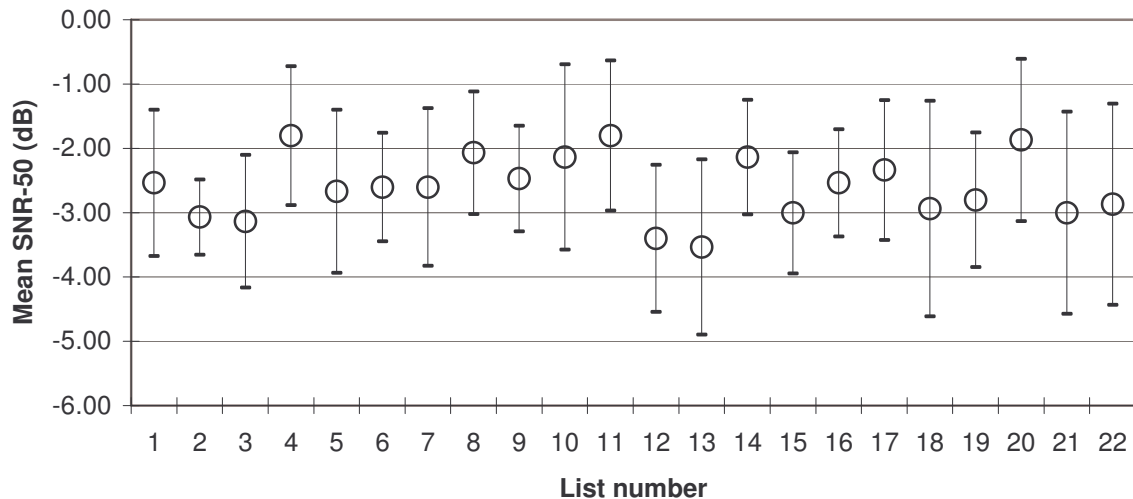


Figure 4.18: Mean SNR-50 across subjects (n=10) for each of the 22 slope lists (unfiltered). Error bars indicate +/- one standard deviation for each list.

The minimum mean as shown in Figure 4.18 was -3.53 (list 13), and the maximum was -1.8 (list 11), indicating that all the means fell within a range of 1.73 dB. The size of the standard deviation varied between 0.59 and 1.68 dB, indicating that all standard deviations were smaller than 2 dB.

From these results, it was also possible to calculate how much the results for each list deviated from the overall mean (across all lists and all subjects). This was determined by comparing each subject's score on each list with the overall mean of all subjects across all lists. The mean and standard deviation of these differences or deviations across subjects could then be calculated for each list. These findings are illustrated in Figure 4.19, and indicate that all the lists deviated less than 1 dB from the overall mean. The standard deviations of these differences ranged between 0.59 and 1.68, thus all below 2 dB.

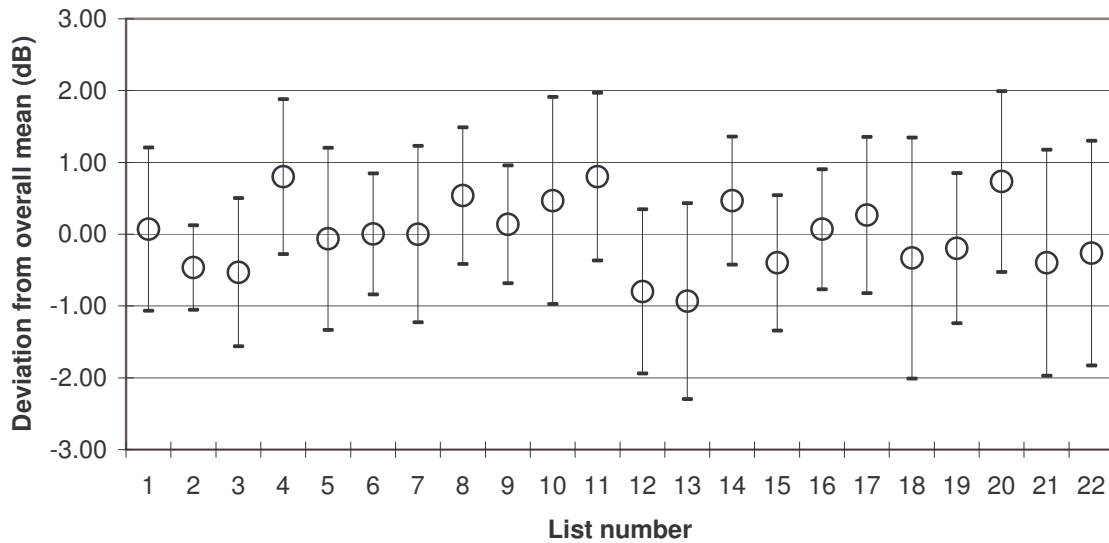


Figure 4.19: Mean differences between the average of each slope list and the overall mean. Error bars indicate +/- one standard deviation from these means.

The values depicted in Figure 4.19 provide an indication of the inter-list equivalence. To further evaluate the equivalence of these lists, a Friedman two-way analysis of variance was conducted to determine whether there were significant differences between the scores of these lists. When the overall difference between all the slope lists was investigated, the Friedman test statistic revealed a p-value of 0.0146, which indicates a significant difference (<0.05). However, multiple comparisons were also conducted to determine whether there were any two lists that differed significantly from each other. This time, no significant differences were found between any two lists.

4.4.2.2 Phonetically balanced lists

As a control procedure for the above-mentioned experiment, the 22 PB lists were subjected to the same procedures as the slope lists, using a different group of subjects that adhered to the same selection criteria. For these lists, a mean SNR-50 of -2.87 dB was obtained across subjects ($n=10$), with a standard deviation of 0.76 dB. The mean SNR-50 threshold and standard deviations of each of the subjects are shown in Table 4.10

Table 4.10: Means and standard deviations of SNR-50 for each subject across the 22 PB lists (unfiltered).

SUBJECT NO.	MEAN	STD DEV
1	-1.94	1.79
2	-2.88	0.96
3	-3.18	1.17
4	-3.42	1.00
5	-1.97	1.21
6	-3.33	1.54
7	-2.97	1.23
8	-3.06	1.05
9	-3.15	1.25
10	-2.76	1.39
Maximum	-1.94	1.79
Minimum	-3.42	0.96
Overall mean	-2.87	1.26

For the PB lists, the poorest mean (indicated by the highest SNR) was -1.94 dB, and the best performance was a mean of -3.42 dB. Within-subject variability is illustrated by the standard deviations as shown in Table 4.10, and ranged between 0.96 and 1.79 dB, therefore all below 1.8 dB.

The mean score and standard deviation for each of the PB lists across subjects ($n=10$) are indicated in Figure 4.20. These means lay between the best performance at -4.13 dB (list 8) and the poorest at -1.4 dB (list 20), a total range of 2.73 dB. Standard deviations for lists ranged between 0.69 and 1.5 dB.

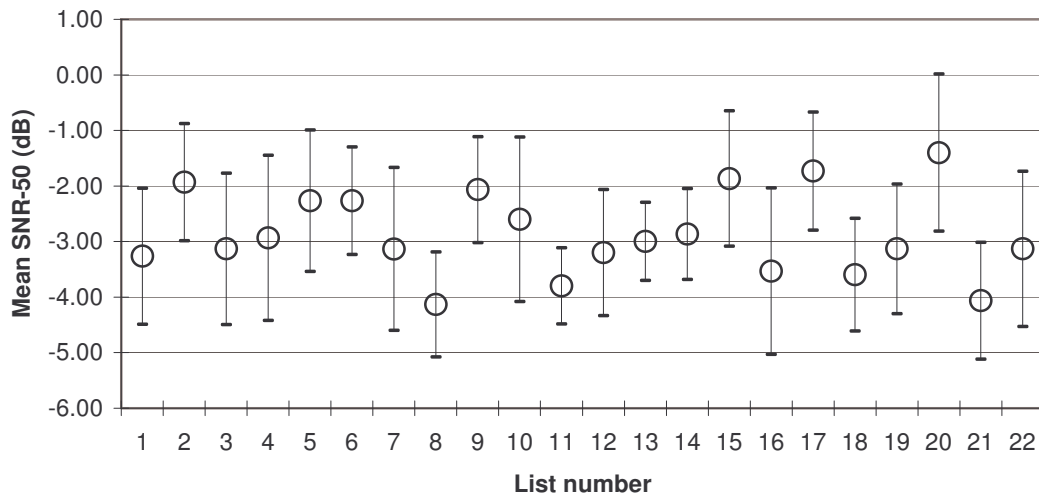


Figure 4.20: Mean SNR-50 across subjects (n=10) for each of the 22 PB lists (unfiltered). Error bars indicate +/- one standard deviation for each list.

The scores of each subject on each list were also compared to the overall mean of all subjects for all lists. Subsequently, the mean deviation from the overall mean could be calculated for each list, along with the standard deviation from this mean. These values are depicted in Figure 4.21. The mean deviations of the PB lists were slightly higher than those of the slope lists, but were all still within +/- 1.5 dB from the overall mean. The standard deviations from these means varied between 0.69 and 1.5 dB.

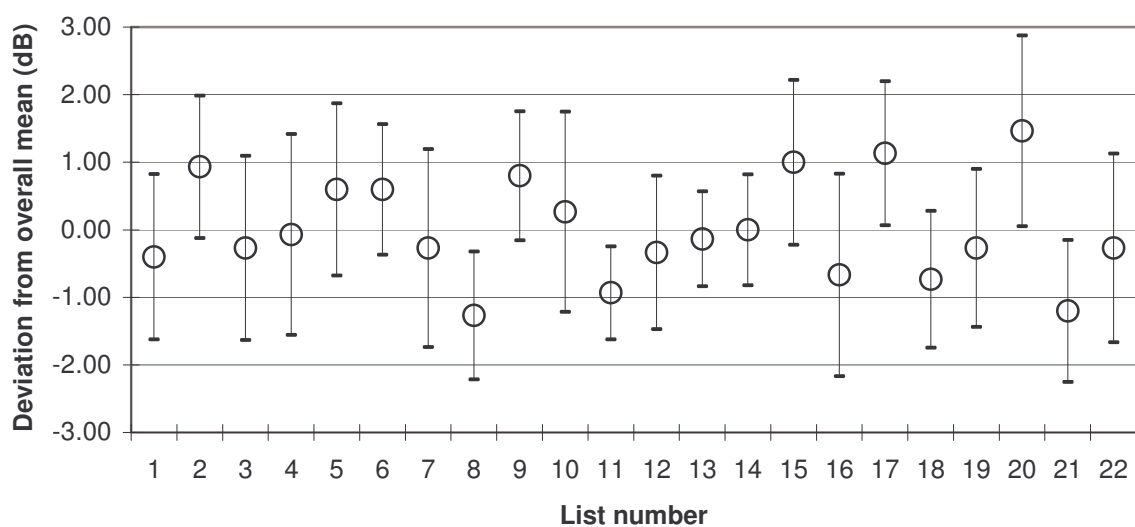


Figure 4.21: Mean differences between the average of each PB list and the overall mean. Error bars indicate +/- one standard deviation from these means.

The variability between lists was further investigated by examining the significance of the differences in scores between lists. The Friedman two-way analysis of variance revealed that there was an overall significant difference (p-value <0.0001) between all the PB lists, i.e. the list used for testing had a significant effect on results. Additionally, the lists were compared in pairs, and several significant differences were found, as shown in Table 4.11.

Table 4.11: Comparison of unfiltered PB list pairs

LIST NO.	DIFFERED SIGNIFICANTLY FROM:
List 8	List 2
	List 9
	List 15
	List 17
	List 20
List 11	List 17
	List 20
List 21	List 2
	List 15
	List 17
	List 20

As shown in the table, lists number 8, 11 and 21 differed significantly from a number of other lists. To determine how the lists with significant differences compared in terms of SNR-50 values, the mean SNR-50 values (across listeners) for the 22 lists were arranged in rising order according to SNR-50 value and are depicted in Figure 4.22.

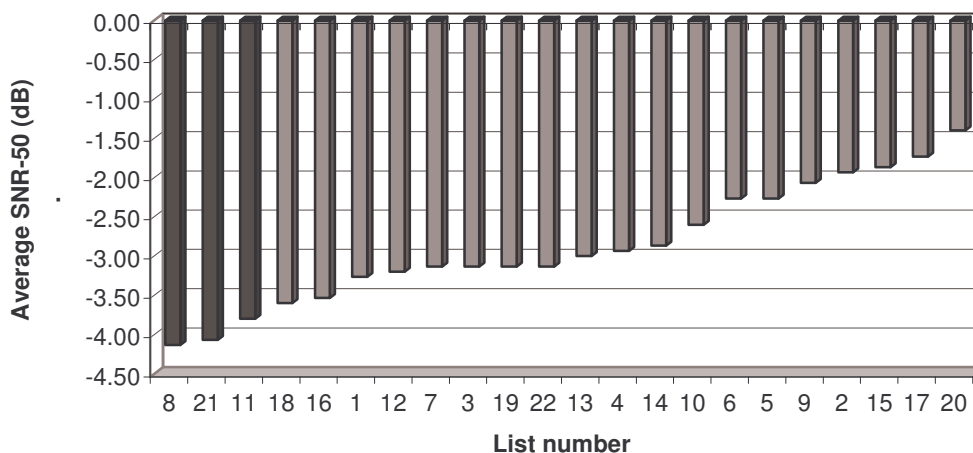


Figure 4.22: Unfiltered PB lists arranged in order of mean SNR-50 scores

From the illustration in Figure 4.22 it is clear that the lists that differed significantly from each other were at the extremes of the arrangement. Lists number 8, 21 and 11 were the lists with the lowest scores (poorest SNR at 50% intelligibility), whereas the lists that differed significantly from these lists (numbers 2, 9, 15, 17, and 20) had the highest scores. In addition, Lists number 8, 11, and 21 showed the largest deviations from the overall mean (-1.27, -0.93 and -1.20 dB respectively). The five highest scoring lists also showed large deviations from the overall mean, ranging from 0.8 to 1.47 dB.

4.4.3 Experiment II: List application in listeners with simulated loss

For the second experiment in this phase, both sets of lists were again presented to the same two groups of subjects, but this time with the sentence material filtered in order to simulate a high frequency hearing loss. Testing for the second experiment was conducted approximately one week after the subject had been tested for the first experiment (minimum waiting time 7 days, maximum 11 days). Each subject listened to the same set of lists, presented in the same order as for the first experiment. However, the sentence material was filtered for this second experiment (using a low-pass filter from 2 kHz with a roll-off slope of 48 dB/octave), thereby simulating a high frequency hearing loss in the listeners. For this reason, lists are consistently called “filtered” in

this section, to avoid confusion with the results attained in the first experiment. The results for the two list sets are presented separately in this section.

4.4.3.1 Slope lists

The mean SNR-50 of the 22 filtered slope lists across subjects (n=10) was (+)0.99 dB, with a standard deviation of 1.05 across subjects. The mean SNR-50 thresholds and standard deviation thereof for each subject across all lists are depicted in Table 4.12.

Table 4.12: Means and standard deviations of SNR-50 for each subject across the 22 filtered⁹ slope lists.

SUBJECT NO.	MEAN	STD DEV
1	0.67	1.79
2	0.15	1.65
3	1.21	1.28
4	1.33	1.83
5	2.09	1.85
6	0.52	1.67
7	1.03	1.52
8	0.85	1.80
9	1.85	2.26
10	0.24	1.99
Maximum	2.09	2.26
Minimum	0.15	1.28
Overall mean	0.99	1.76

As indicated in Table 4.12, the maximum value (poorest score, as indicated by the better SNR level required for 50% performance) was 2.09 dB. The minimum value (best performance) was at 0.15 dB, resulting in a performance range of 1.94 dB across subjects. The within-subject variability is indicated by the standard deviation in performance for each subject. These standard deviations are also shown in Table 4.12, indicating that the largest deviation within a subject was below 2.3 dB.

⁹ The term “filtered” used in this section indicates that the material was filtered to simulate a high frequency hearing loss in the listeners.

The mean score and standard deviation for each of the filtered slope lists across subjects ($n=10$) are indicated in Figure 4.26. These means lay between the best performance at -0.87 dB (list 3) and the poorest at 3 dB (list 15), a total range of 3.87 dB. Standard deviations for lists ranged between 0.63 and 2.72 dB.

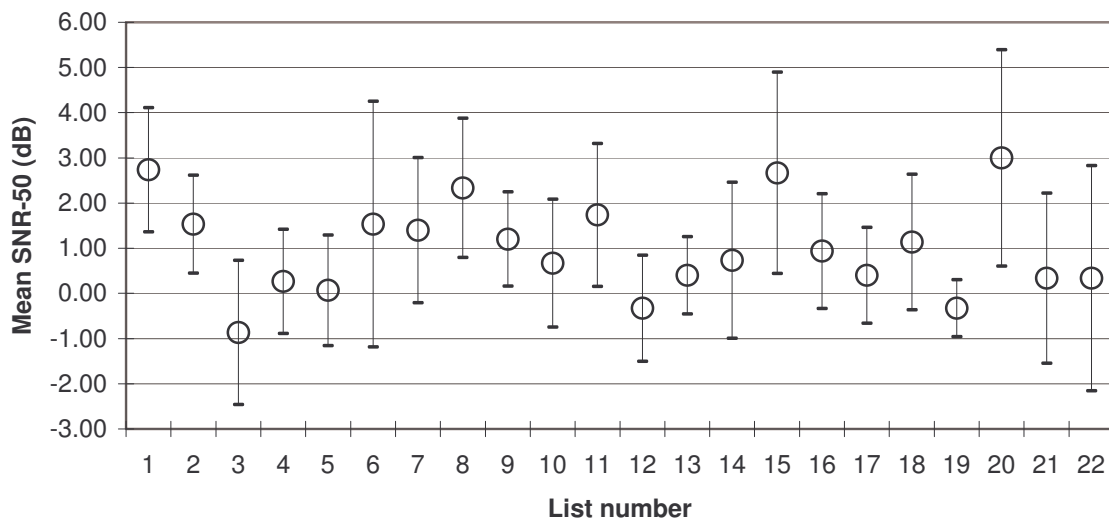


Figure 4.23: Mean SNR-50 across subjects ($n=10$) for each of the 22 filtered slope lists. Error bars indicate \pm one standard deviation for each list

The scores of each subject on each list were also compared to the overall mean of all subjects for all lists. Subsequently, the mean deviation from the overall mean could be calculated for each list, along with the standard deviation from this mean. These values are depicted in Figure 4.24.

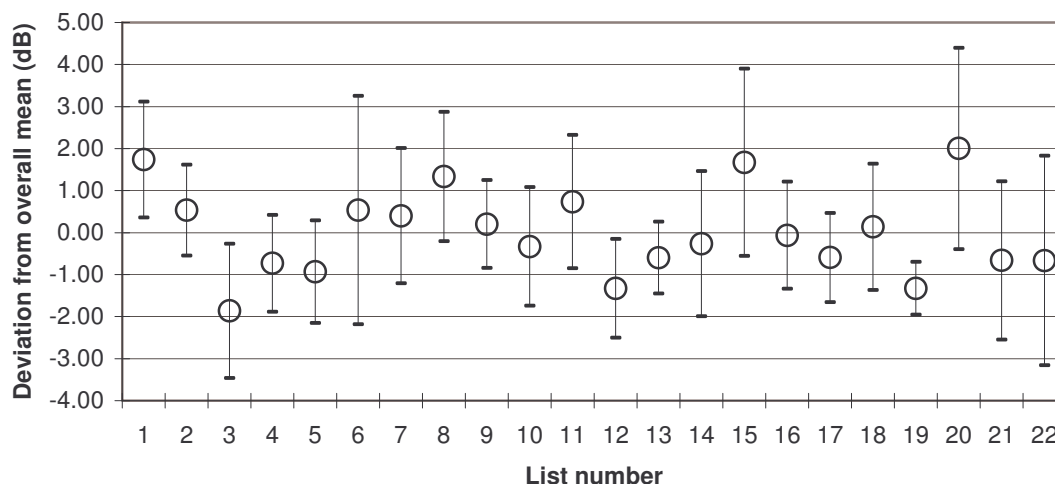


Figure 4.24: Mean differences between the average of each filtered slope list and the overall mean. Error bars indicate +/- one standard deviation from these means.

These results provide an indication of the variability between the individual lists. In addition, an analysis of variance was also conducted to determine the significance of the differences between lists. The results of this statistical procedure revealed a p-value < 0.0001, indicating that there was a significant effect of list number on performance. Subsequently, individual lists were compared in pairs and several significant differences were found, as indicated below in Table 4.13.

Table 4.13: Comparison of filtered slope list pairs

LIST NO.	DIFFERED SIGNIFICANTLY FROM:
List 3	List 1 List 8 List 15 List 20
List 12	List 1 List 20
List 19	List 1 List 8 List 15 List 20

As indicated in the table, Lists 3, 12, and 19 were the lists that differed significantly from a number of other lists, namely Lists 1,8, 15 and 20. In order to compare the scores of the lists that differed significantly, all the SNR-50 means were arranged in order, and are depicted in Figure 4.25.

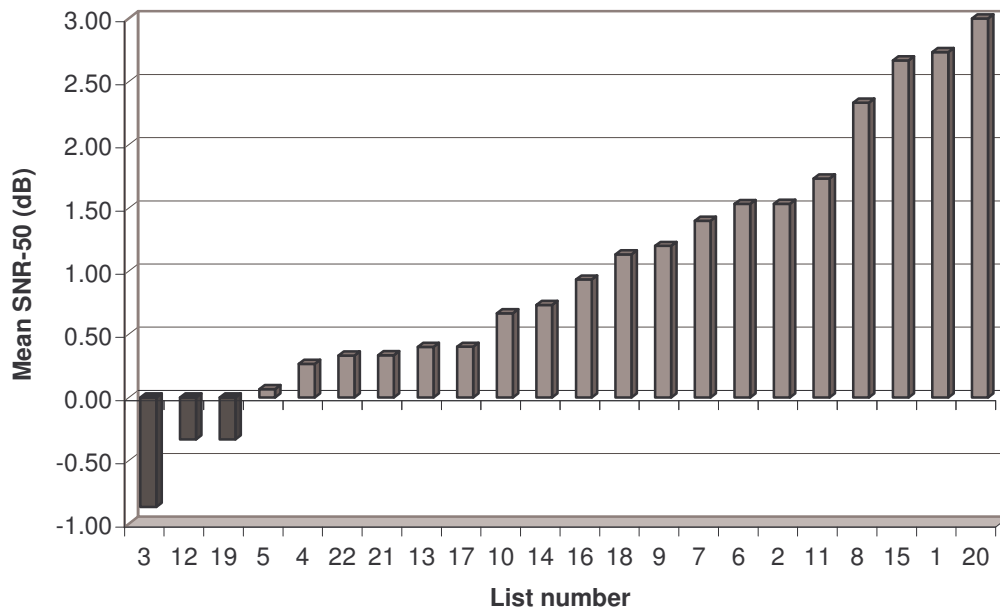


Figure 4.25: Filtered slope lists arranged in order of mean SNR-50 scores

The graph illustrates that Lists 3, 12, and 19 yielded the lowest means (best performance), whereas the lists that differed significantly from these (1, 8, 15, and 20) required a much better SNR to yield a 50% performance. In addition, Lists 3, 12, and 19 each showed a large deviation from the overall mean (>1 dB, as seen in Figure 4.24). List number 20, which elicited the poorest performance or highest SNR, had the largest deviation from the mean (+2.01), and also differed significantly from the three lists shown in Table 4.13.

4.4.3.2 Phonetically balanced lists

The mean SNR-50 of the 22 PB lists across subjects ($n=10$) was (+) 0.81 dB, with a standard deviation of 0.92 across all subjects. The mean SNR-50 thresholds and standard deviation thereof for each subject across all lists are depicted in Table 4.14.

Table 4.14: Means and standard deviations of SNR-50 for each subject across the 22 filtered PB lists.

SUBJECT NO.	MEAN	STD DEV
1	2.52	1.64
2	1.06	1.52
3	0.76	1.91
4	0.39	2.00
5	0.39	1.73
6	0.94	2.11
7	-0.58	1.61
8	1.15	1.45
9	0.48	1.20
10	0.94	1.70
Maximum	2.52	2.11
Minimum	-0.58	1.20
Overall mean	0.81	1.69

As shown in Table 4.14, subjects' means ranged from -0.58 dB to 2.52 dB (a range of 3.09 dB). Intra-subject variability in performance as indicated by the standard deviations were on average 1.69 dB, and ranged between 1.2 and 2.11, thus also below 2.5 dB.

The mean score and standard deviation for each of the filtered PB lists across subjects ($n=10$) are indicated in Figure 4.26. These means lay between the best performance at -1.20 dB (list 21) and the poorest at 2.73 dB (list 15), a total range of 3.93 dB. Standard deviations for lists ranged between 1.06 and 2.42 dB.

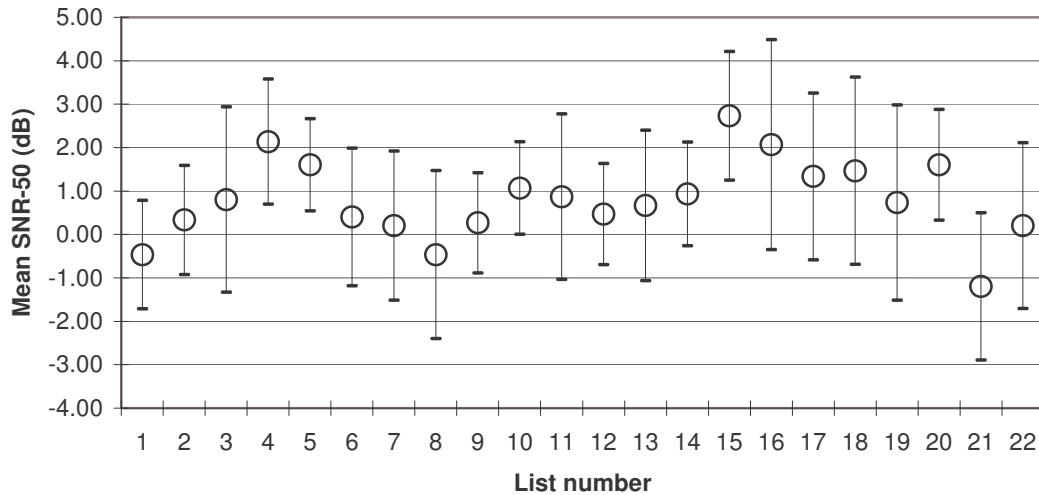


Figure 4.26: Mean SNR-50 across subjects (n=10) for each of the 22 filtered PB lists. Error bars indicate +/- one standard deviation for each list.

Using these means, it was possible to compare the scores for each list with the overall mean (mean score for all lists across all subjects). Each list's deviation from the overall mean is indicated in Figure 4.27 below.

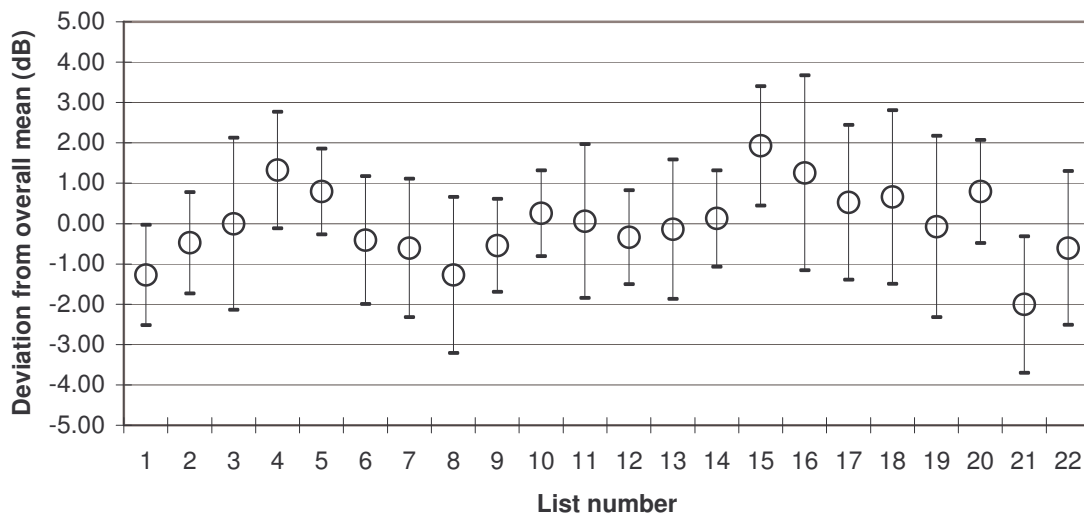


Figure 4.27: Mean differences between the average of each filtered PB list and the overall mean. Error bars indicate +/- one standard deviation from these means.

The differences between the filtered PB lists were also investigated. The analysis of variance indicated that there were significant differences within the collection of lists ($p < 0.0001$), and paired comparisons revealed which lists differed significantly from each other. These findings are indicated in Table 4.15.

Table 4.15: Comparison of filtered PB list pairs

LIST NO.	DIFFERED SIGNIFICANTLY FROM:
List 15	List 1 List 8 List 21
List 21	List 4 List 5 List 15 List 16

The table indicates that Lists number 15 and 21 differed significantly from a number of other lists. The lists' results were subsequently arranged according to SNR-50 means, and this arrangement is depicted in Figure 4.28.

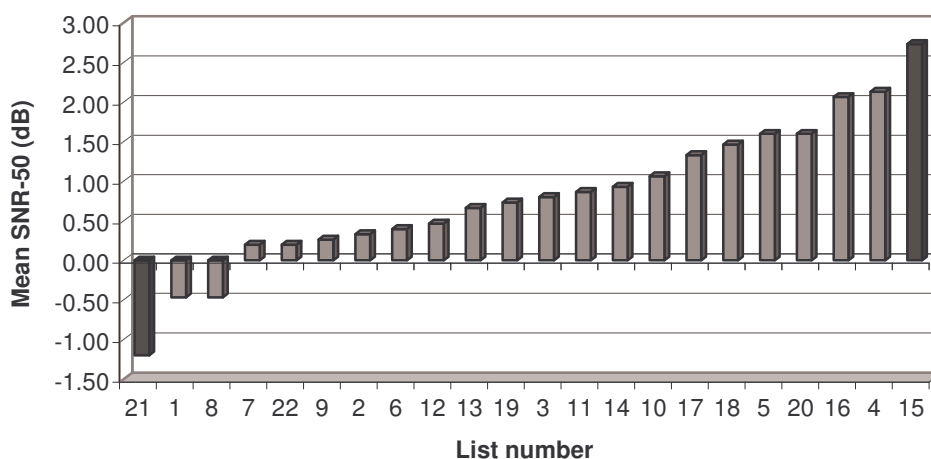


Figure 4.28: Filtered PB lists arranged in order of mean SNR-50 scores

Figure 4.28 indicates that the two lists that differed significantly from a number of other lists were at the extremes of the arrangement, with List number 21

yielding the best performance (poorest SNR required for 50% intelligibility), and List number 15 showed the poorest performance. Lists 15 and 21 also showed the largest deviations from the overall mean in this experiment (1.93 and -2.01 dB respectively).

4.4.4 Comparison of results from Experiment I and II

The results of the two experiments conducted during this phase shed some light on the performance of the two sets of lists. This section will provide a concise overview of these findings by comparing the results of the two experiments and the experimental and control conditions' results. Table 4.16 below illustrates these comparisons.

Table 4.16: Comparison of results from two experiments

VARIABLES		SLOPE LISTS (dB)	PB LISTS (dB)	DIFFERENCE (dB)
Unfiltered (Exp I)	Mean*	-2.60	-2.87	
	Standard deviation*	0.50	0.76	0.27
	Minimum [†]	-3.53	-4.13	
	Maximum [†]	-1.80	-1.40	
	Range [†]	1.73	2.73	1.00
	Within-subject std dev	1.14	1.26	0.12
Filtered (Exp II)	Mean	0.99	0.81	
	Standard deviation	1.05	0.92	-0.12
	Minimum	-0.87	-1.20	
	Maximum	3.00	2.73	
	Range	3.87	3.93	0.06
	Within-subject std dev	1.76	1.69	-0.07
Difference (Exp I and II compared)	Mean	3.60	3.67	
	Standard deviation	0.97	0.90	
	Minimum	2.07	2.27	
	Maximum	5.67	5.60	
	Range	3.60	3.33	
	Wilcoxon p-value	<0.05	<0.05	

* Means and standard deviations are for all lists across all subjects

[†] Minimum refers to the list that scored the lowest mean across participants, whereas the maximum indicates the highest mean scored. "Range" specifies the difference between this minimum and maximum.

The table provides a clear comparison of the results from both experiments. The right-hand column elucidates the differences in the deviations and ranges of the two sets, which indicates the variability within each set. Positive values indicate that the PB lists showed greater variability, whereas negative values denote greater variability in the slope lists. In the section referring to the differences between unfiltered and filtered lists, the results for the Wilcoxon-rank sum test are indicated. This procedure was used to determine the magnitude and direction of differences for pairs of scores, and can be used to test the significance of the difference between dependent samples (Maxwell and Satake, 2006:339). The results provided an indication of the significance of the differences between the unfiltered and filtered conditions for both sets of lists. Each individual list's scores for the first and second experiment were compared. It was found that each of the lists in both sets showed significant differences ($p < 0.05$) between filtered (simulated loss) and unfiltered (normal-hearing) conditions. This measure could provide an indication of the sensitivity of the lists to the presence of a peripheral hearing impairment, since the filtered results represent a simulated hearing loss.

4.5 Conclusion

The three phases of the research process each yielded its own set of results. Many of these results were required to facilitate the initiation of the following phase, but there were also findings that served to enhance the existing body of knowledge on the development of a test for sentence recognition in noise. Therefore, the data collected for this study not only provides insight into the reliability and characteristics of the test developed here, but also indicates the efficacy of the methodology followed in the development of such a test. This information could assist future researchers in the development of similar tests, and will be discussed in the following chapter.

4.6 Summary

Chapter 4 presented the results from each of the three research phases. The findings of the first phase were mainly related to the development of the sentence material, whereas the results from second phase revealed the intelligibility of these sentences in the presence of noise, thereby facilitating the elimination of lists to improve the uniformity of the collection. The third phase described the results of the list compilation process, as well as the experimental application of these lists in normal-hearing subjects as well as listeners with a simulated hearing loss. The results depicted in this chapter are discussed and compared with relevant literature in the next chapter.

5. DISCUSSION

5.1 Introduction

The main aim of the study was to develop a valid and reliable Afrikaans test of sentence recognition thresholds in noise. To attain this aim, three phases of research were conducted. The current chapter provides a comprehensive discussion of all these results. The findings and methods of the research are critically discussed and compared to previous findings and methods reported in the literature.

5.2 Phase I: Compiling and refining test materials

The aim of the first phase of the research was to develop a collection of recorded Afrikaans sentences suitable for the assessment of speech recognition in noise. The sequence of procedures followed to achieve this aim is discussed in this section by referring to relevant literature and critically evaluating the results and procedures of the current research.

5.2.1 Compilation of sentences

A collection of written sentences was compiled from three different sources – the BKB sentences (Bench and Bamford, 1979), the “Afrikaanse Reseptiewe Woordeskattoets” (ARW, compiled by Buitendag in 1994), and the “Foneties gebalanseerde Woordelyste” for children 3-5 years or PBC lists (Tesner and Laubscher, unpublished). The sources for the sentence material are compared to the sources used by previous studies of a similar kind in Table 5.1. The table shows that most previous researchers have either created their own material, or translated and adapted existing material. The current study used a combination of these two methods, as was also done by Kollmeier and Wesselkamp (1997).

Table 5.1: Sources used for sentence material by current and previous studies

AUTHORS	SOURCES FOR SENTENCE MATERIAL
Plomp and Mimpfen (1979)	Own sentences
Nilsson et al (1994)	BKB sentences
Kollmeier and Wesselkamp (1997)	Existing German sentence test plus own material
Versfeld et al (2000)	Newspapers
Vaillancourt et al (2005)	6-7 year old reading level material
Wong and Soli (2005)	HINT and original material
Hällgren et al (2006)	HINT translated and adapted
Van Wieringen and Wouters (2006)	Original material
Wong et al (2007) (Mainland Mandarin)	Original material
Wong et al (2007) (Taiwanese Mandarin)	Original material
Current study	Translated BKB plus own material from word lists

The majority of sentences for the current collection were translations of the BKB sentences, and a smaller percentage of the collection consisted of original sentences compiled from the vocabulary of the ARW and the PBC lists. The method of translating existing material therefore provided the largest number of sentences and also served as an example of the typical sentence structure and content to be followed in the composition of the original material. The combination of these two methods was successful in producing a large collection of sentences that all adhered to the style and content criteria stipulated in the literature. These criteria are reviewed in Table 5.2, showing that the sentences compiled in the current research were complete sentences (containing a verb and a noun), representative of everyday speech, and free from proverbs, questions, exclamations and proper nouns. These characteristics ensure that sentences are not too redundant or confusing (Plomp and Mimpfen, 1979:44) and correlate well with those of previously developed sentence collections, as demonstrated in Table 5.2.

Table 5.2: Characteristics/criteria of sentences used in previous and current research

AUTHORS	EVERYDAY / CONVERSATIONAL	COMPLETE SENTENCES (VERB & NOUN)	NO PROVERBS, QUESTIONS, EXCLAMATIONS, PROPER NOUNS
Plomp and Mimpfen (1979)	✓		✓
Nilsson et al (1994)		✓	
Kollmeier and Wesselkamp (1997)		✓	
Versfeld et al (2000)	✓	✓	✓
Vaillancourt et al (2005)			✓
Wong and Soli (2005)	✓		
Hällgren et al (2006)	✓		
Van Wieringen and Wouters (2006)	✓	✓	✓
Wong et al (2007) (Mainland Mandarin)	✓		
Wong et al (2007) (Taiwanese Mandarin)	✓		
Current study	✓	✓	✓

In addition to the style and content characteristics, other researchers have given consideration to the length of the sentence materials. The sentence lengths reported by previous researchers, along with the length of the current sentences, are depicted in Table 5.3.

Table 5.3: Sentence length reported by previous and current researchers

AUTHORS	SENTENCE LENGTH
Plomp and Mimpfen (1979)	8-9 syllables
Kollmeier and Wesselkamp (1997)	3-7 words
Nilsson et al (1994)	6-7 syllables
Hällgren et al (2006)	5-9 syllables
Vaillancourt et al (2005)	5-7 syllables
Wong and Soli (2005)	4-7 characters
Van Wieringen and Wouters (2006)	4-15 syllables
Wong et al (2007) (Taiwanese Mandarin)	10 characters (1 word = 1-3 characters)
Wong et al (2007) (Mainland Mandarin)	10 characters (1 word = 1-3 characters)
Versfeld et al (2000)	8-9 syllables
Current study	3-9 syllables

Table 5.3 demonstrates that many other studies showed smaller variability in terms of sentence length than the current study. Differences in the amount of syllables may be due to inherent differences in word length between languages (Hällgren et al., 2006:228). However, the variation in terms of number of words (4 to 8 words, or a range of 5 words) was identical to that of Kollmeier and Wesselkamp (1997:2413) and the range of syllables (3 to 9, implying a range of 7 syllables) was smaller than that of Van Wieringen and Wouters (2006, Figure 1), which ranged from 4 to 15 syllables. Therefore, the uniformity in length of the current sentence collection can be considered similar to that of previously compiled sentence collections. This is an important consideration to ensure homogeneity of the collection.

5.2.2 Rating of naturalness

As shown in Table 5.2, many previous researchers have claimed that their sentences were representative of everyday or conversational speech. However, not all of these researchers have assessed whether or not their sentences were actually considered to be natural or conversational by native speakers of the language. Some researchers used experts in speech therapy and audiology to select sentences that were representative of conversational speech (Versfeld et al., 2000:1672).

Other researchers did assess the naturalness of the material by having native speakers of the language in which the test was developed assess this aspect (Nilsson et al., 1994:1086; Vaillancourt et al., 2005:360; Hällgren, 2006:228; Wong et al., 2007:71S). This method ensures that members of the general population that use the language find the material understandable. Vaillancourt et al. (2005:360) further specified that the participants who rated the naturalness had to be from different age groups, educational and geographical backgrounds to ensure that the material is considered natural by all members of the language group, regardless of their educational history or geographical background. The current study also followed this approach.

As with previously reported research (Nilsson et al., 1994:1086; Vaillancourt et al., 2005:360; Hällgren, 2006:228; Wong and Soli, 2005:279; Wong et al., 2007:71S), naturalness of the current sentences was rated on a 7-point scale (1=artificial; 7=natural) by a number of native speakers of the relevant language (Nilsson et al., 1994:1086; Vaillancourt et al., 2005:360; Hällgren, 2006:228; Wong and Soli, 2005:279; Wong et al., 2007:71S). Although some researchers have reported rejecting sentences with a rating lower than five and then proceeding to record the material (Wong et al., 2007:71S), other researchers have adapted sentences with a mean rating less than 6 and had these sentences rated by a second group of participants (Nilsson et al., 1994:1086; Vaillancourt et al., 2005:360; Hällgren, 2006:228). The latter model was followed by the current study.

Nilsson et al. (1994) did not specify the number of sentences that were submitted to a second round of rating, but stated that during the second round all the revised sentences received a naturalness rating of 6 or above. Vaillancourt et al. (2005:360) reported that 165 of their original 524 sentences were altered after the first round of rating, and 4 sentences were eliminated. Hällgren et al. (2006:228) reportedly revised 79 of their sentences, and eliminated 4 sentences after the second round. These findings are illustrated and compared with the current study in Table 5.4.

Table 5.4: Results of naturalness rating of previous and current research

AUTHORS	NO. OF SENTENCES REVISED & RE-RATED	NO. OF SENTENCES EXCLUDED FROM COLLECTION
Nilsson et al (1994)	Not specified	Not specified
Vaillancourt et al (2005)	165	4
Hällgren et al (2006)	79	4
Current study	12	3

Table 5.4 indicates that the current study had a much smaller number of sentences altered after the first round in comparison to previous studies, possibly indicating that sentences had a higher degree of naturalness to begin

with. However, after the second rating, other studies showed a similar number of eliminations (4) than the current study (3). The methodology used in the current study to evaluate naturalness was therefore comparable to that of previous researchers. The small number of sentences that required alteration after the first rating may be related to the criteria used for the compilation of the sentences, where the objective was to compile sentences that are complete, (containing a verb) (Versfeld et al., 2000:1672), representative of everyday speech, and free from proverbs, questions, exclamations and proper nouns (Plomp and Mimpen, 1979:44). Table 5.2 illustrated how the current study and previous studies compared in terms of these criteria. Since following these criteria led to the compilation of a sentence collection that received such a high naturalness rating, future researchers may find it useful to adhere to these criteria when compiling sentence material of a similar nature.

5.2.3 Rating of grammatical complexity

In addition to rating the naturalness of the sentences, the grammatical and syntactic complexity of the present sentences was also analysed. The majority (82%) of the sentences were found to be on a three-year old level, and the remaining 18% fell within the four- to five-year old classification. Other researchers have stated that their sentences were simple and/or understandable for children six and up (Nilsson et al., 1994:1086; Wong and Soli, 2005:278; Wong et al., 2007:71S), but have not reported assessment of the grammatical complexity of the material.

Nilsson et al. (1994:1090) have used commercial software to determine the readability of the sentences, and described this as a simple, repeatable and objective technique for estimating grammatical and syntactic level. The assumption of this method, though, is that readability correlates directly with syntactic and grammatical complexity. Most readability formulas consider only word difficulty, sentence length and some index of sentence difficulty (DiStefano and Valencia, 1980:247). However, these measures may not

ensure that reading material is of the same level of difficulty, as it does not consider syntactical complexity, which has been found to influence reading comprehension, and therefore possibly the auditory comprehension of material (DiStefano and Valencia, 1980:250). Furthermore, commercial software grading the readability of material may be readily available for the English language, but this is not the case for many other languages, and other methods of determining grammatical and syntactic complexity must be followed for such languages.

Due to the fact that previous researchers have not conducted a grammatical rating similar to the present study, current results cannot be compared to previous findings. However, the data describing the grammatical level of the sentences have provided some interesting insights when compared to the results of the second phase. These will be discussed in section 5.3, along with the other findings of the second phase.

5.2.4 Recording and editing of material

Due to the significant differences between individual speakers that have been reported in the literature (Versfeld et al., 2000:1675; Wilson et al., 1990:777), it is important that pre-recorded material be used in the development of a speech-in-noise test. Therefore, the current sentence collection was recorded digitally in a sound-proof booth, following the rating of naturalness and grammatical complexity. A female speaker was used for the recordings, as was done by Plomp and Mimpen (1979:44) and Hällgren et al. (2006:228) in the development of similar tests. The effect that gender or individual differences could have on the intensity level of the material, was mitigated by digitally recording and adjusting the intensity level of the speech material (Wilson and Strouse, 1999:1337; Nilsson et al., 1994:1087). Although gender difference does not always imply a bigger difference between two speakers (Versfeld et al., 2000:1676), individual differences between speakers can be generally accepted (Wilson et al., 1990:774), and results acquired using test material from a test presented by a specific speaker should therefore be

cautiously compared to results obtained with a different speaker, regardless of the gender of that speaker.

To avoid the effect that dialect might have on the test results (Lyregaard, 1997:49), a panel of judges assisted in selecting the speaker used for the recordings to ensure that her speech was free of dialectical influences. This was in accordance with methods followed by previous researchers (Versfeld et al., 2000:1672; Vaillancourt et al., 2005:360) and proved to be an effective manner of ensuring standard pronunciation, as a random sample of sentences from the recorded collection was judged by an expert in speech science in phonetics to be of a standard variety.

It should be noted here that the listeners used in the experiments described in Phase II and III were all residents from the same area and therefore are currently exposed to the same regional accent on a daily basis. Should the current recordings be used on individuals from a different geographic area who may be exposed to a different (non-standard) regional accent, it may be necessary to obtain norms within such a population before making clinical deductions from the findings.

5.3 Phase II: Selecting an equivalent subset of sentences

The second phase of the research was aimed at selecting from the total collection of 515 sentences those sentences that are equally intelligible in the presence of noise. This entailed two equalisation procedures. During the first procedure, one fixed SNR was used and sentences showing similar performance at this condition were selected. The second procedure involved the selection of sentences that presented with similar intelligibility slopes (performance as a function of SNR), as determined by their performance at three different SNRs. The results for these two procedures are discussed separately in this section.

During the first equalisation procedure, all sentences ($n=515$) were presented to all subjects ($n=10$) at a fixed SNR of -5 dB. This SNR was based on results of previous studies, as discussed in Chapter 2 (Table 2.4), and was anticipated to yield an intelligibility score of 50%. The results correlated well with the anticipated score, as the overall mean across participants and sentences was 52%. The standard deviation from this mean was 27%, which agreed with the standard deviation of 27.1% found by Hällgren et al. (2005:229) during their initial experiment. Sentences that did not fall within one standard deviation from the mean were excluded from the collection, and this resulted in a collection of 330 sentences. Other researchers have also excluded or adapted sentences that did not fall within 20-25% of the mean (Vaillancourt et al., 2006:361; Wong et al., 2007:71S).

A second equalisation procedure was conducted where the 330 sentences remaining after the first procedure of this phase were presented to a new group of subjects at two different SNRs (-2 and -8 dB). This enabled the researcher to determine the intelligibility slopes of these sentences by plotting the performance between these two points and the data point determined by the previous procedure (at SNR-5). The mean difference between performances at -2 and -8 dB was 71%, indicating a slope of 11.83%/dB. In Table 5.5 this value is compared to results found by other researchers that used sentence material in speech-weighted noise.

Table 5.5: Intelligibility slopes of previous and current studies

AUTHORS	SLOPE (%/dB)
Plomp & Mimpen (1979)	15
Versfeld et al (2000)	11.1 - 12.5
Wong & Soli (2005)	9.7
Vaillancourt et al (2005)	10.3
Hällgren et al (2006)	15.4 - 17.9
Van Wieringen & Wouters (2006)	17.5
Wong et al (2007)	9
Mean	12.9
Current study	11.8

As shown in the table, other studies of a similar nature have yielded slopes in the range of 9 to 17.9%/dB, and the 11.83% slope of the current study compares well to the slopes reported by other studies. To efficiently measure a threshold of speech recognition in noise, a steep psychometric function is necessary (Versfeld et al., 2000:1676). This slope indicates to what extent the SNR determines the intelligibility of the sentence (Plomp and Mimpen, 1979:49). The 11.83 %/dB slope indicates that small changes in SNR yielded notable changes in speech recognition, increasing the precision with which differences in speech recognition can be detected. This slope is also much steeper than the 2-3 %/dB slope that has been reported for word recognition in quiet (Wilson and Carter, 2001:12), demonstrating that the type of test developed in this study is a more sensitive method of detecting small differences in speech recognition than, for example, traditional spondee threshold tests.

Previous studies have used these slopes as an indication of the SNR adjustments needed to equate the intelligibility of the sentences (Vaillancourt et al., 2005:361; Wong and Soli, 2005:281; Wong et al., 2007:71S). In the present study, however, the slopes were used to identify sentences with a similar performance under different conditions, which enabled the researcher to exclude sentences that differed significantly from the majority, much like Versfeld et al. (2000:1678). Excluding sentences instead of re-scaling intensities reduced the number of subjects and hours of data collection, as previous researchers have sometimes had to conduct up to seven rounds of testing in order to verify the effect of the re-scaling procedure (Nilsson et al., 1994:1088). This meant that a large number of the originally recorded sentences had to be excluded, and only 222 of the original 515 sentences, or 43% of the sentences remained. This ratio is compared to other studies that have mainly followed the re-scaling method in Table 5.6.

Table 5.6: Percentage of original sentences retained in previous and current studies

AUTHORS	RE-SCALED / ELIMINATED	STARTED WITH	ENDED WITH	% OF ORIGINAL SENTENCES
Plomp & Mimpen (1979)	Both	170	130	76
Nilsson et al (1994)	Re-scaled	336	252	75
Kollmeier & Wesselkamp (1997)	Re-scaled	324	200	62
Versfeld et al (2000) Speaker 1	Eliminated	1272	726	57
Versfeld et al (2000) Speaker 2	Eliminated	1272	783	62
Wong & Soli (2005)	Re-scaled	549	240	44
Vaillancourt et al (2005)	Eliminated	520	240	46
Hällgren et al (2006)	Re-scaled	313	279	89
Van Wieringen & Wouters (2006)	Eliminated	730	355	49
Wong et al (2007) Mainland	Re-scaled	816	240	29
Wong et al (2007) Taiwanese	Re-scaled	752	240	32
Mean		641	335	56
Current study		515	222	43

The table illustrates that other studies have managed to retain between 29 and 89% of their original collection, and the study at hand retained 43%. On average, other studies managed to retain approximately 56% of initial sentences despite the fact that many of these studies have re-scaled the intensities of sentences and conducted several rounds of testing in order to retain as many sentences as possible. It appears therefore that the current method of excluding sentences instead of re-scaling intensities may be equally efficient in selecting a large collection of sentences, without the need for several groups of subjects and many rounds of testing. Should there be need for a larger collection of sentences in the future, it may be possible to increase the size of the current collection by re-scaling some of the rejected sentences according to these slopes and evaluating their equivalence in normal-hearing subjects.

The question remains, however, how well the equivalence of the current set of sentences compare to those of other studies. In order to make this comparison, the results obtained for the final 222 sentences selected during

Phase II had to be converted to dB-values. This is because the present study worked mainly with the percentage intelligibility scores and excluded sentences based on certain criteria, whereas other studies re-scaled intensities of sentences and were able to calculate the mean intensity and standard deviation of sentences at the 50% mark.

To facilitate comparison with the results of other studies, the slope found during the second phase (11.83%/dB) was used to convert the results attained for each sentence at SNR-5 (closest to 50%) to the dB level where that sentence would attain 50%. These findings are compared to similar studies reporting on this aspect in Table 5.7.

Table 5.7: Mean SNR at 50% with deviations shown for current and previous studies¹⁰

AUTHORS	MEAN SNR	STD DEV	% SENTENCES WITHIN +/- 1dB	% SENTENCES WITHIN +/- 2dB
Plomp and Mimpen (1979)	x	1.3 dB	x	x
Nilsson et al (1994)	-5.3	x	51%	x
Kollmeier & Wesselkamp (1997)	-6.1	0.94 dB	x	x
Vaillancourt et al (2005)	-5	x	47%	x
Wong & Soli (2005)	-6	x	x	80%
Wong et al (2007)	-5	x	x	80%
Current study	-5.2	1.3	44%	87%

As demonstrated in Table 5.7, the current study yielded a similar mean SNR at 50% recognition than previous studies of a similar nature. The current findings also compared well to previous research in terms of the deviations from this mean. The standard deviation of the current findings (1.3 dB) is identical to that reported by Plomp and Mimpen (1979:44), and only slightly larger than the 0.94 dB reported by Kollmeier and Wesselkamp (1997:2414). A slightly smaller percentage of the current sentences (44%) fell within +/- 1

¹⁰ Some similar studies referred to elsewhere in the report are not listed in the table, as the format in which those results were reported did not allow for a direct comparison to the current findings. Also, not all studies expressed the deviation from the mean in the same manner – some reported the standard deviation, whereas others only specified the percentage of sentences that fell within 1 or 2 dB from the overall mean, hence the “x” symbols in the table.

dB of the mean than the 51% reported by Nilsson et al (1994:1088), but this percentage differed only 3% from the 47% reported by Vaillancourt et al. (2005:361). In addition, the current study showed a relatively high percentage of sentences that fell within +/- 2 dB of the overall mean, namely 87% compared to the 80% findings of both Wong and Soli (2005:282) and Wong et al. (2007:71S). Therefore, the current method managed to yield a collection of sentences with a similar degree of equivalence to previous studies, without the need for several rounds of retesting, deeming it both effective and time efficient.

Additional findings made during the second phase included the effect of gender, the correlation between grammar rating and intelligibility, and the practice effect. During the present research, a significant difference ($p < 0.0001$) between male and female performance was found during the first equalisation procedure of the second phase. None of the other similar studies found in existing literature have mentioned any gender effects, and in fact did not mention the gender composition of their research samples. This effect was therefore re-evaluated during the second equalisation procedure, but this time the effect was not significant. It is not clear from the results why this gender effect occurred during the first procedure, but did not repeat during the second. It is possible that the increased uniformity of the sentence collection used during the second procedure led to less variability between subjects, which eliminated the gender effect found during the first experiment when there were large differences in intelligibility of sentences in the collection. Due to the uncertainty surrounding the effect of gender, the subject samples in the following phase of the project were composed of an equal number of males and females to prevent gender from having an influence on the findings.

The influence of gender on speech recognition in noise is not widely reported in the literature. Neuro-imaging studies have reported conflicting findings on gender differences in terms of brain activation patterns during certain auditory tasks, but differences do appear to exist in terms of noise perception and

auditory working memory (Ruytjens et al., 2007:2074). It has recently been found that even simple sounds (such as music or noise) induce different brain activation patterns in different gender groups, especially in the primary auditory cortex (Ruytjens et al., 2007:2079). In view of these reports, and the inconclusiveness of the present findings on this matter, further research using larger sample groups and specifically focusing on this issue is warranted. Similar studies aimed at the development of a test using speech material in noise should also pay heed to this variable when selecting a research sample.

The grammar rating awarded to each sentence during the first phase was also compared to its intelligibility during the second phase, another factor not reported by previous studies. The current findings indicated no significant correlation between grammatical complexity and intelligibility in noise. This may be due to the fact that the sentence collection was relatively uniform in terms of its grammatical complexity (all sentences were between three- and five-year old level). Should a sentence collection contain a greater variety in terms of grammatical complexity, it may be worth investigating this factor, following the methods described in this study.

The effect of practice on performance was also assessed during the second phase by comparing the performance of subjects within the first 20 sentences presented to them. During the first equalisation procedure (SNR-5), mean performance within the first 20 sentences ranged between 5% and 44%, a range of 39%. However, a Friedman two-way analysis of variance (ANOVA) test resulted in a p -value > 0.05 for the first ten, second ten and first twenty sentences, indicating that the order of presentation was insignificant. Despite the fact that no significant practice effect was demonstrated, two practice lists were used during the second equalisation procedure in an attempt to reduce the range between scores. During this second procedure, a smaller range was found within the first 20 sentences (33% and 20% for SNR-2 and SNR-8 respectively) and there were again no significant differences in performance on the first 20 sentences. It is possible, however, that the smaller range was

due to the ceiling or floor effect at these intensities, since the maximum score at SNR-2 was 100% and the minimum at SNR-8 was 0%.

Plomp and Mimpen (1979:48) have investigated the practice effect over 10 lists consisting of 13 sentences each and have found an improvement in performance during the course of the test. Other researchers did not report on the effect of practice, but did present a practice list or two prior to testing (Nilsson et al., 1994:1088; Versfeld et al., 2000:1677; Wong and Soli, 2005:283; Wong et al., 2007:72S). Although the current findings did not show significant effects in terms of practice, the possibility of this effect should not be excluded, and the commonly reported method of using practice lists is perhaps the safest way of excluding the possible effect of this variable.

5.4 Phase III: Compilation and evaluation of lists

The aim of the final phase of the research project was to compare the inter-list reliability and response variability of two list sets compiled using two different methods of list compilation. Results were obtained for the compilation process, experimental application in normal-hearing subjects, as well as for application in listeners with a simulated high frequency hearing loss. The results of each of these processes will be discussed in this section.

5.4.1 List compilation

During list compilation, the experimental method where lists were arranged according to the intelligibility slopes of the sentences was found to be very time efficient, as only 14 exchanges were needed to obtain a high degree of equivalence between lists. In addition, the lists that were compiled in this manner were predicted (according to the findings of the second phase) to be highly equivalent when used with individuals with normal peripheral hearing, with a range of 6% at various SNRs. The PB lists required a great deal of effort and time to compile as described in the methodology of Phase III, and required the use of specialised mathematical knowledge and software. Also,

when calculating the intelligibility slopes of these lists according to previous rounds of testing, it could be predicted that these lists would be less equivalent in noise when presented to normal-hearing subjects than the slope lists. This was because at each data point on the slope (SNR-8, SNR-5 and SNR-2), there were greater differences in the scores of the PB lists than there were between the slope lists. The differences between the list with the maximum value and the list with the minimum value at each SNR for each set of lists are illustrated in Table 5.8. The table clearly shows that the slope lists were more equivalent at each SNR, whereas the PB lists showed great variability in its performance at each point.

Table 5.8: Difference in list equivalence between two list sets, demonstrated by differences (in percentage) between best and worst scoring lists at each SNR.

LIST TYPE	RANGE AT SNR-8	RANGE AT SNR-5	RANGE AT SNR-2
PB lists	12%	18%	8%
Slope lists	6%	6%	6%
Difference	6%	12%	2%

In terms of phonetic content, the PB lists naturally outperformed the slope lists. The difference in errors on the total phoneme counts, as well as the errors per list for the PB and slope lists are compared in Table 5.9. Other studies that have reported on the precision of their phonetic balance are also included in the table. The table shows that the PB lists in the current study outperformed not only the slope lists, but also lists of previous studies in terms of the total phoneme counts that were within +/- 1 of an “ideal count”. The procedure used to compile PB lists in the current study (as described in the methodology of Phase III) therefore proved to be an accurate technique to accomplish phonetic balance, and provides a useful guideline for future researchers developing similar tests.

Table 5.9: Comparison of phonetic balance of current and previous studies

VARIABLES	PB LISTS	SLOPE LISTS	NILSSON ET AL (1994)	VAILLANCOURT ET AL (2005)	WONG & SOLI (2005)	HÄLLGREN ET AL (2006)
% phoneme counts where error = 0	35%	26%	30-35%	35%	<15%	x
% of counts within +/- 1 phoneme	83%	67%	68%	75%	+/- 61%	70%
Largest error on single count	+5	-12	x	x	x	x
Number of errors ≥ +/- 5	1 (<0.1%)	38 (3.5%)	x	x	x	x
Minimum errors per list	8	21	x	x	x	x
Maximum errors per list	18	37	x	x	x	x
Mean number of errors per list	12	29	x	x	x	x

As demonstrated in the table, even the slope lists of the current study, that were not compiled according to phonetic content, compared well with other studies that did pay heed to phonetic content. Despite this, there were still a large number of phonetic errors in some of the slope lists (on average 29 errors per list), and one list had an error count on one phoneme of -12 (an excess of 12). These errors are likely to have an effect if the material is applied to a population that find some phonemes more difficult to hear than others, such as individuals with a high frequency hearing loss.

5.4.1.1 Experiment I: Application to normal-hearing listeners

In order to determine and compare the equivalence of the two list sets, both sets were applied to a group of normal-hearing subjects, and re-tested with the same group using a low-pass filter to simulate a hearing loss. This was done in two separate experiments, where Experiment I represented the tests conducted on the normal-hearing subjects, and Experiment II tested the same subjects, but this time with a filtered version of the same sentence material.

A number of other studies have applied their lists to normal-hearing subjects and have reported on the repeatability, reliability, and list equivalence of the sentence material. The findings reported by these studies, along with the findings of Experiment I of the current study are shown in Table 5.10.

Table 5.10: List equivalence of current and previous sentence lists

AUTHORS	MEAN SNR50 ACROSS SUBJECTS (dB)	STD DEV FROM MEAN (dB)	WITHIN- SUBJECT STD DEV	DEV FROM OVERALL MEAN (dB)
Plomp & Mimpen (1979)	x	x	0.9	x
Nilsson et al (1994)	-2.9	0.78	1.13	All within +/- 1dB
Kollmeier & Wesselkamp (1997)	-6.2	0.27	x	x
Versfeld et al (2000)	-4.1	0.56	1.07	x
Vaillancourt et al (2005)	-3.3	0.5	1.1	All within +/- 1dB
Wong & Soli (2005)	-3.9	1	1.8	All within +/- 1dB
Hällgren et al (2006)	-3	1.1	x	All within +/- 1dB
Van Wieringen & Wouters (2006)	-7.8	1.2	1.17	All within +/- 1dB
Wong et al (2007) MHINT-M	-4.3	0.62	0.89	x
Wong et al (2007) MHINT-T	-4	0.94	0.75	x
Overall mean	-4.39	0.77	1.10	
Slope lists	-2.6	0.50 dB	1.14 dB	All within +/- 1dB
PB lists	-2.9	0.76 dB	1.26 dB	All within +/- 1.5dB

It should be noted that many of the other studies have experimented with either the number of sentences per list (e.g. Wong and Soli 2005:283) or the directionality of the noise (e.g. Vaillancourt et al., 2005:363), and the results shown in the table are the findings made under conditions that showed greatest similarity to the current study (i.e. using 10-sentence lists; and speech and noise coming from the same direction or “noise front” condition). The mean SNR-50 across subjects (represented in the first column) indicates the mean threshold obtained with all lists or all sentences across the entire group of subjects. As shown in the table, means found by other studies ranged from -7.8 to -2.9 dB, with an overall mean of -4.39 . The current study’s lists resulted in SNR-50 means of -2.6 and -2.9 dB for slope and PB

lists respectively, which corresponded closely with the findings of Nilsson et al. (1994:1090), but was higher than many of the other studies, who reported thresholds as low as -7.8 (Van Wieringen and Wouters, 2006:9). This indicates that the material in the current study required a better SNR for 50% recognition, and may therefore be more difficult than material from previous reports.

The reason for the difference between the means of the current study and those reported by previous researchers is not clear. It is interesting to note, however, that the study showing the closest correspondence to the current findings in terms of mean SNR-50 (Nilsson et al., 1994:1090) was also the only other study that used binaural headphone presentation of the test material. Other studies have used monaural presentation (Kollmeier and Wesselkamp, 1997:2413; Versfeld et al., 2000:1674; Van Wieringen and Wouters, 2006:5), sound-field presentation (Hällgren et al., 2006:229) or simulated sound-field conditions (Vaillancourt et al., 2005:363; Wong and Soli, 2005:283). Previous research has shown that word recognition scores are normally better with binaural free-field presentation than in a monaural free-field condition where one ear is occluded (Feuerstein, 1992:82; Persson, Harder, Arlinger and Magnuson, 2001:629). The fact that the current study and the HINT study (Nilsson et al., 1994) using binaural presentation have required better SNRs than other studies using monaural presentation indicates that either the nature of the speech material, or the use of headphones versus free-field presentation affects the difference between monaural and binaural listening. Further investigation is needed to explore this discrepancy and to determine whether the material used in the current research will yield different results if presented using a different auditory transmission channel.

Perhaps more important than the absolute values of the thresholds are the standard deviations of these means, which give an indication of the variability within the collection of lists. Other studies have reported standard deviations

ranging from 0.27 to 1.2 dB (with a mean of 0.77 dB). The current study found a standard deviation of 0.5 dB for the slope lists, and a slightly larger (but still comparable to other findings) 0.76 dB for the PB lists, indicating a high degree of equivalence across lists.

The within-subject standard deviations of the current study, as represented in the table, were calculated by determining the standard deviation for each subject across all lists, and then calculating the mean of this standard deviation across subjects. For the slope lists, this deviation was within 0.01 dB of that reported by Nilsson et al. (1994:1090). The deviation for the PB lists was slightly larger (1.26 dB), indicating a greater degree of variability between lists for specific subjects. A large deviation for a particular subject would mean that different lists would yield largely different threshold values for that particular subject. By implication, if a patient is re-tested with a different list to evaluate, for instance, the effect of a change in hearing aid settings, differences in scores may be purely due to differences in lists, and not due to the actual hearing aid settings. For the present study, standard deviations for the slope lists ranged between 0.9 and 1.45 dB, i.e. the maximum standard deviation for a particular subject was less than 1.5 dB, indicating a high degree of equivalence. The PB lists showed slightly larger deviations, ranging from 0.96 to 1.79. However, this implies a standard deviation that is still smaller than 2 dB.

The range between minimum and maximum scores attained by each subject across all lists was also determined and is depicted in Table 5.11 below. The table shows the best and worst scores of each subject across lists. This indicates the range of variability in results that could be expected when testing one individual with different lists.

Table 5.11: Within-subject score ranges

SLOPE LISTS											
SUBJECT NO.:	1	2	3	4	5	6	7	8	9	10	MEAN
Maximum	-0.33	-1.00	-1.67	-0.33	-0.33	-0.33	-0.33	-1.00	0.33	-0.33	
Minimum	-5.00	-5.67	-5.00	-4.33	-5.67	-3.67	-4.33	-5.67	-4.33	-5.67	
Range	4.67	4.67	3.33	4.00	5.34	3.34	4.00	4.67	4.66	5.34	4.40
PB LISTS											
SUBJECT NO.:	1	2	3	4	5	6	7	8	9	10	MEAN
Maximum	1.00	-1.00	-1.00	-1.67	0.33	1.00	-0.33	-1.00	-1.00	0.33	
Minimum	-5.00	-4.33	-5.67	-5.00	-4.33	-5.67	-6.33	-5.00	-5.00	-5.00	
Range	6.00	3.33	4.67	3.33	4.66	6.67	6.00	4.00	4.00	5.33	4.80

As shown in the table, the ranges for the slope lists were generally smaller than those of the PB lists, with the largest range in one subject being 5.34 dB, whereas the PB lists yielded ranges of up to 6.67 dB. This means that for a single subject, there could be a difference of up to 6.67 dB in scores attained with different lists. To avoid the effect that this variability could have on test results, it may be wise to use more than one list at a time when assessing sentence recognition in noise, and determining the SNR-50 threshold by calculating the mean score across two or more lists.

The deviation from the overall mean was also calculated for each list. In the current study, the average deviations of all the slope lists were found to be within +/- 1 dB from the overall mean. This finding corresponded with a number of other studies (Nilsson et al., 1994:1090; Vaillancourt et al., 2005:363; Wong and Soli, 2005:285; Hällgren et al., 2006:231; Van Wieringen and Wouters, 2006:9). The deviation from the overall mean was slightly larger for the PB lists (ranging between -1.27 and 1.47), but still within 1.5 dB from the overall mean.

Scores attained with each of the lists were also compared to the scores of other lists to determine if there were significant differences between individual lists. As described in the previous chapter, there were no significant

differences between any two of the slope lists. There were, however, a number of PB lists that differed significantly from each other. Lists number 8, 11 and 21 were the lists with the lowest SNR-50 scores and these lists differed significantly from the lists that yielded the highest scores (Lists 2, 9, 15, 17, 20). To determine the reason for the significant differences between these lists, the number of errors in the phonetic balance of these lists was investigated. Table 5.12 below shows all the PB lists and the number of errors in phonetic balance for each, arranged in ascending order.

Table 5.12: Number of errors in phonetic balance for PB lists, arranged in ascending order

LIST NO.	NO. OF ERRORS IN PHONETIC BALANCE
13	8
22	8
1	9
2	9
14	9
15	9
7	10
6	11
16	11
18	11
4	12
10	12
20	12
3	13
11	13
17	13
8	14
12	14
19	14
21	14
5	16
9	18
Mean no. of errors ("problematic lists")	12.8
Mean no. of errors (other lists)	11.3

As shown in the table, the mean number of errors for lists that differed significantly from other lists was slightly higher than the mean of the other lists. However, not all of these lists had a high number of errors. List number

2, for example, had only 9 errors in phonetic balance (the fourth lowest number of errors), but still differed significantly from a number of other lists. List number 5, in turn, had a high number of errors (16 in total, the second highest number of errors), but did not differ significantly from any other lists. It appears therefore that the number of errors in exact phonetic balance do not fully account for the differences between lists.

The reason for these differences, along with the larger deviations (standard deviations, deviation from overall mean, and within-subject deviations) of the PB lists is not clear, especially since the accuracy of the current study's phonetic balancing appeared to be greater than other studies. Other factors that may have influenced list equivalence which was considered by other researchers include the number of syllables per list (Van Wieringen and Wouters, 2006:6), the number of words, and the number of phonemes per list (Kollmeier and Wesselkamp, 1997:2414). It is possible that the slope lists happened to contain sentences of greater equivalence in terms of these characteristics, purely because the sentences were grouped into lists based on their known difficulty. When considering all the deviations and differences, it is clear that the slope lists compared very well with previous studies in terms of repeatability and list equivalence, and although the PB lists were less equivalent, the deviations were still not dissimilar to those reported in the literature.

5.4.1.2 Experiment II: Application in listeners with simulated hearing loss

Not many studies have reported on the equivalence of lists when applied to a hearing-impaired population, or in the presence of a simulated hearing loss. Van Wieringen and Wouters (2006:12) have applied their sentence material to a number of adult cochlear implantees to assess the ability of the material to successfully map the hearing abilities of both good and poor performers. It was found that the sentence material was indeed suitable for this population, but list equivalence was not compared.

Nilsson et al. (1994:1093) conducted a bandwidth study to determine the effect of audible bandwidth on the reliability of the measurements. This was motivated by an anticipation that sentence thresholds may sometimes be measured under conditions of reduced bandwidth, either due to a hearing impairment or limited bandwidth transmission systems. Different reductions in bandwidth were tested and it was found that thresholds increased significantly in reduced bandwidth conditions. In addition, there were significant differences between thresholds under these conditions (Nilsson et al., 1994:1094). However, standard deviations did not increase significantly when the 4kHz octave band was eliminated, but only when the 2kHz octave band was also eliminated (Nilsson et al., 1994:1095). With this bandwidth, standard deviations for 5 lists in noise increased to about 3 dB.

In the current research, a filter was used to simulate a hearing loss from 2kHz upwards (with a roll-off slope of 48 dB/octave), and the same groups of subjects used for the initial experimental application of the lists were re-tested under this condition. The results of this experiment showed an increase in standard deviation for both the slope lists and PB lists, although the deviations were still well below the deviation of 3 dB reported by Nilsson et al. (1994:1095). For the slope lists, the standard deviation of all lists across subjects increased from 0.5 dB in the unfiltered condition (Experiment I) to 1.05 dB in the filtered experiment. The standard deviation of the PB lists increased from 0.76 dB to 0.92 dB. Both these sets therefore still showed standard deviations close to 1 dB, despite the additional variability between lists caused by the filter.

Within-subject standard deviations also increased for both sets of lists. For the slope lists, the mean within-subject standard deviation increased from 1.14 dB in the unfiltered condition to 1.76 dB in the filtered condition, a difference of 0.62 dB. The deviation of the PB lists increased from 1.26 dB (unfiltered) to 1.69 dB (filtered) – a difference of 0.43 dB. Although both these increases

were still quite small (only about half a dB), the reason for the increased amount of variation in subjects' scores was explored by comparing individual lists in pairs.

It was found that in both list sets, there were pairs of lists that differed significantly from each other. Within the set of slope lists, there were mainly 3 lists (numbers 3, 12, and 19) that differed significantly from other lists (1, 8, 15, and 20). The number of errors in phonetic balance within these lists was explored in an attempt to account for the differences between them. These values are shown in Table 5.13 below.

Table 5.13: Number of errors in phonetic balance for slope lists, arranged in ascending order

LIST NO.	NO. OF ERRORS IN PHONETIC BALANCE
13	21
19	24
1	26
2	26
21	26
4	27
3	28
5	28
6	28
7	29
9	29
15	29
16	29
17	29
11	30
14	30
22	30
10	31
20	31
12	32
18	33
8	37
Mean no. of errors ("problematic lists")	29.6
Mean no. of errors (other lists)	28.4

The findings demonstrated in this table reveals that on average the “problematic lists” (those that differed significantly from a number of other lists) had a slightly higher number of errors in phonetic balance than lists that did not differ significantly from other lists. Despite this, there were lists with a relatively low number of errors (such as List 19) that differed significantly from others, and lists with a large number of errors (such as List 18) that did not differ significantly from any of the other lists. Once again it appears that phonetic content alone cannot account for differences between lists, especially since the filter affected particular frequency areas. If large errors in phonetic balance meant that a particular list did not have the same phonetic content (and therefore frequency composition) as a number of other lists, one would have expected this list to differ significantly from other lists if a frequency specific filter is applied. This was, however, not the case for a number of slope lists (e.g. Lists 18, 10, and 22).

In the filtered condition, a number of PB lists also yielded SNR-50 scores that were significantly different when compared in pairs. List 15 (the highest scoring list) differed significantly from Lists 1, 8, and 21 (the lowest scoring lists), whereas List 21 differed significantly from 4, 5, 15, and 16 – all of which attained relatively high scores. Referring back to Table 5.12, these lists did not have the highest number of errors in phonetic balance. List 15, in fact, had a very small number of phonetic errors (9 in total). Again it seems that phonetic content alone does not account for differences between lists.

In light of all these results, the experimental (slope) and control (PB) lists yielded results that were very similar under both the unfiltered and filtered conditions. In the first experiment (unfiltered), however, the slope lists outperformed the PB lists in terms of list equivalence, seeing as the standard deviation, the range between minimum and maximum as well as within-subject standard deviations were smaller for these lists. Looking at the comparisons between pairs of lists, the slope lists also performed better than the PB lists in the first experiment, since no two slope lists differed

significantly from each other, whereas three of the PB lists showed significant differences with two or more lists.

In the filtered condition (Experiment II), the PB lists showed a slightly smaller standard deviation (0.92) than the slope lists (1.05). However, the range between the minimum and maximum was smaller for the slope lists than for the PB lists, and both sets presented with significant differences between lists (overall as well as in paired comparisons). As far as the difference between SNR-50 scores in the filtered and unfiltered conditions is concerned, the results of the two sets were nearly identical.

To improve uniformity between lists, two different options were considered. The first option was to eliminate from the collection lists that differed significantly from other lists, thereby reducing the size of the collection. A second possibility was to divide the lists into two groups, where the lists in each group would be equivalent. This option would imply that patients that have been tested by a list in a specific group should always be tested with other lists from the same group if results are to be compared.

The option of elimination was first applied to the slope lists. Since there were no significant differences between any two of these lists in the unfiltered condition, the results from the filtered condition were used to determine which lists should be excluded. As demonstrated in the previous chapter, lists number 3, 12 and 19 differed significantly from a number of other lists. These lists also showed large deviations from the overall mean (>1 dB). With these three lists eliminated from the collection, there would be no significant differences between any two lists in the collection. List number 20 was also excluded, since it had the highest average score in the filtered condition and a large deviation from the overall mean in both experiments (filtered and unfiltered). The effect that these eliminations had on the results is demonstrated in Table 5.14 below.

Table 5.14: Comparison of results for slope lists, with and without problematic lists

	VARIABLES	ALL SLOPE LISTS	SLOPE LISTS EXCLUDING LIST 3,12,19,20	DIFFERENCE
UNFILTERED (EXPERIMENT I)	Mean	-2.60	-2.56	
	Standard deviation	0.50	0.46	-0.04
	Minimum	-3.53	-3.53	
	Maximum	-1.80	-1.80	
	Range	1.73	1.73	0.00
	Within-subject standard deviation	1.14	1.14	0.00
	Deviation from overall mean	All within +/- 1dB	All within +/- 1dB	
FILTERED (EXPERIMENT II)	Mean	0.99	1.13	
	Standard deviation	1.05	0.83	-0.22
	Minimum	-0.87	0.07	
	Maximum	3.00	2.74	
	Range	3.87	2.67	-1.20
	Within-subject standard deviation	1.76	1.62	-0.14
	Deviation from overall mean	All within +/- 2dB	All within +/- 1.75	

Table 5.14 demonstrates that the elimination of lists number 3, 12, 19 and 20 removed significant differences between any two lists in the collection. In addition, the standard deviation for the unfiltered slope lists was slightly reduced (from 0.5 to 0.46 dB), and the standard deviation for the filtered slope lists was reduced even more (from 1.05 to 0.83 dB). Furthermore, the range between the minimum and maximum (mean scores) was reduced in the filtered condition from 3.87 dB to below 3 dB (2.67). This method therefore seemed quite effective in improving the equivalence of the slope lists, although it entailed reducing the number of lists from 22 to 18. A higher degree of list equivalence implies a higher degree of reliability, and the exclusion of these lists therefore improved the performance of the test.

The second option was also applied to the slope lists by arranging the sentences into two groups according to performance in the filtered condition

(seeing as this was the only condition where significant differences between specific lists were observed). By doing this, the pairs of lists that differed significantly from each other were allocated to different groups, and therefore no significant differences occurred within a group. The means for each group under each experimental condition, along with the standard deviations, minimum, maximum and range (difference between minimum and maximum) are depicted in Table 5.15.

Table 5.15: Values calculated separately for Groups 1 and 2* of the slope lists, compared to values for all 22 lists

	VARIABLES	GROUP 1	GROUP 2	ALL LISTS
UNFILTERED (EXP I)	Mean	-2.71	-2.50	-2.60
	Standard deviation	0.56	0.43	0.50
	Minimum	-3.53	-3.07	-3.53
	Maximum	-1.80	-1.80	-1.80
	Range	1.73	1.27	1.73
FILTERED (EXP II)	Mean	0.15	1.84	0.81
	Standard deviation	0.48	0.72	0.92
	Minimum	-0.87	0.93	2.73
	Maximum	0.73	3.00	-1.20
	Range	1.60	2.07	3.93

* Group 1 = Lists 3,4, 5, 10, 12, 13, 14, 17, 19, 21, 22

Group 2 = Lists 1, 2, 6, 7, 8, 9, 11, 15, 16, 18, 20

The table shows that in the unfiltered condition, the standard deviation across lists was not improved for Group 1 (in fact it increased by 0.06 dB) and was only slightly reduced (by 0.07 dB) for Group 2. The range between minimum and maximum scores was also reduced only for one of the groups (Group 2). In the filtered condition, however, the standard deviation was reduced with 0.4 and 0.2 dB for Groups 1 and 2 respectively, and the range between minimum and maximum scores was reduced for both groups (with 2.33 and 1.86 respectively).

The positive effect of dividing the slope lists into two groups therefore extended mainly to the filtered condition. However, since the slope lists already showed such a high degree of equivalence in the unfiltered condition and improvement was mainly needed for the filtered condition, the method of grouping lists may be a valid method of improving list equivalence. Unfortunately, this method would reduce the number of lists that have comparable results to 11 (half the original number). For this reason, the option of eliminating “problematic” lists as discussed earlier, may be a more efficient manner of improving the equivalence of lists without reducing the size of the collection significantly.

Subsequently, the option of elimination was applied to the PB lists. This proved to be more complex than with the slope lists, since the lists that showed significant differences from each other were not the same for the filtered and unfiltered conditions. In the unfiltered condition, lists number 8, 11 and 21 differed significantly from a number of other lists and their elimination would mean that there would be no two lists that differed significantly. However, for the filtered condition, lists 8 and 21 were again problematic, but this time list number 11 did not differ significantly from any other lists and was in fact very close to the overall mean (mean deviation from the overall mean for list 11 was 0.06). List 15, which did not differ from any other lists in the unfiltered condition, showed significant differences from three other lists (1, 8 and 21) in the filtered condition. In this condition, list 15 also showed a large deviation (1.93 dB) from the overall mean. Therefore, a minimum of four lists (lists 8, 11, 15, and 21) had to be eliminated to remove significant differences from the collection under both conditions. The effect of eliminating these lists is demonstrated in Table 5.16.

Table 5.16: Comparison of results for PB lists, with and without problematic lists

	VARIABLES	ALL PB LISTS	PB LISTS EXCLUDING LIST 8,11,15,21	DIFFERENCE
UNFILTERED (EXPERIMENT I)	Mean	-2.87	-2.73	
	Standard deviation	0.76	0.64	-0.12
	Minimum	-4.13	-3.60	
	Maximum	-1.40	-1.40	
	Range	2.73	2.20	-0.53
	Within-subject standard deviation	1.26	1.22	-0.04
	Deviation from overall mean	All within +/- 1.5dB	All within +/- 1.5dB	
FILTERED (EXPERIMENT II)	Mean	0.81	0.88	
	Standard deviation	0.92	0.71	-0.21
	Minimum	-1.20	-0.47	
	Maximum	2.73	2.13	
	Range	3.93	2.60	-1.33
	Within-subject standard deviation	1.69	1.59	-0.10
	Deviation from overall mean	All within +/- 2dB	All within +/- 1.4dB	

As demonstrated in the table, eliminating lists 8, 11, 15, and 21 from the collection led to a slight reduction in standard deviation for both unfiltered and filtered conditions (a reduction of 0.12 dB and 0.21 dB, respectively). The range between minimum and maximum scores was also slightly reduced for both conditions (0.53 dB for unfiltered and 1.33 for filtered condition). Within-subject standard deviations improved only very slightly in the unfiltered condition (with 0.04 dB), and improvement for the filtered condition was also quite small (0.1 dB). Eliminating these lists did reduce the deviation from the overall mean in the unfiltered condition, where all the lists now fell within +/- 1.4 dB from the overall mean (as opposed to +/-2 dB with the problematic lists included). The elimination of these 4 lists therefore improved the equivalence of the PB lists slightly.

An attempt was also made to arrange the PB lists into two groups where there would be no significant differences within either group. However, due to the fact that the scores on the lists and the ranking of the lists according to these scores differed considerably between filtered and unfiltered conditions (see Figure 4.22 and Figure 4.28), it was not possible to arrange the lists into two groups where both the unfiltered and filtered conditions would then yield no significant differences between lists in a group.

A final comparison was made between the two sets of lists, where “problematic” lists were excluded from each set and the results for each set from each of the two experiments were compared. This comparison is depicted in Table 5.17. The last column in the table specifies the difference between the two sets of lists. The table indicates that the slope lists outperformed the PB lists in the unfiltered condition in terms of list equivalence. This is demonstrated by a smaller standard deviation, a smaller range between minimum and maximum scores, a smaller within-subject standard deviation as well as a smaller deviation from the overall mean. These differences are, however, very small (ranging between 0.08 and 0.5 dB). In the filtered condition (second experiment), however, the PB lists were slightly more equivalent than the slope lists, as indicated by the findings in Table 5.17. As with the first experiment, these differences were quite small. The data in the table also indicate that the difference between the two list sets is larger in the unfiltered condition (1.23) where the slope lists perform better, than in the filtered condition (0.57), where the PB lists showed superior equivalence. It therefore seems that, on the whole, the slope lists showed better performance in terms of list equivalence than the PB lists.

Table 5.17: Comparison of slope lists and PB lists with problematic lists excluded

	VARIABLES	SLOPE LISTS EXCLUDING LIST 3,12,19,20	PB LISTS EXCLUDING LIST 8,11,15,21	DIFFERENCE*
UNFILTERED (EXPERIMENT I)	Mean	-2.56	-2.73	
	Standard deviation	0.46	0.64	0.18
	Minimum	-3.53	-3.60	
	Maximum	-1.80	-1.40	
	Range	1.73	2.20	0.47
	Within-subject standard deviation	1.14	1.22	0.08
	Deviation from overall mean	All within +/- 1dB	All within +/- 1.5dB	0.5
TOTAL DIFFERENCE:				1.23
FILTERED (EXPERIMENT II)	Mean	1.13	0.88	
	Standard deviation	0.83	0.71	-0.12
	Minimum	0.07	-0.47	
	Maximum	2.74	2.13	
	Range	2.67	2.60	-0.07
	Within-subject standard deviation	1.62	1.59	-0.03
	Deviation from overall mean	All within +/- 1.75	All within +/- 1.4dB	-0.35
TOTAL DIFFERENCE				0.57

* A positive value in this column indicates that the result of the PB lists had a greater value than the slope lists, whereas a negative value indicates that the result for the PB lists had a smaller value than the slope lists' result.

A Friedman two-way analysis of variance confirmed these results. This non-parametric test serves to detect significant differences between different groups or repeated measures (Maxwell and Satake, 2006:352). A significant effect of list on test results is indicated by a p-value smaller than 0.05. For the slope lists (excluding lists 3, 12, 19, and 20) in the unfiltered condition this value was just below 0.5 (0.04), but in the filtered condition the effect was more significant (p-value = 0.0009). With the PB lists (excluding lists 8, 11, 15 and 20), the effect of list on test results was significant in the unfiltered condition (p-value = 0.0006). In the filtered condition, this effect was not as strong, although still statistically significant (p = 0.02).

For listeners with normal peripheral hearing, the slope lists therefore show greatest promise in terms of a high degree of list equivalence, even without eliminating any lists from the collection. The phonetically balanced set showed a greater degree of variability between lists for listeners with normal hearing, but performed slightly better than the slope lists in the second experiment where a high frequency hearing loss was simulated. The question remains, however, what the effect of a different filter (affecting different frequencies) would have on both these sets. Excluding lists that differed significantly from other lists in the collection improved the performance of both sets and eliminated significant differences within each collection. However, a different filter may reveal significant differences between other lists in the collection.

The fact that the PB lists were not completely equivalent in either the unfiltered or filtered experiments demonstrates that equivalence in terms of phonetic content does not guarantee list equivalence. List equivalence in a filtered condition, such as in the presence of a hearing loss, may be greatly affected despite the fact that lists are phonetically balanced. It is therefore essential to evaluate list equivalence in populations with a hearing impairment or at least a simulated hearing loss before assuming that phonetically balanced lists can be used to reliably evaluate hearing-impaired individuals.

It is possible that a greater degree of list equivalence for both normal-hearing and hearing-impaired individuals could be attained if the two reported methods of list compilation are combined. This could be achieved with the use of an optimisation algorithm, where different parameters are assigned a weighting factor to determine the priority of each of the parameters to be considered (Kollmeier and Wesselkamp, 1997:2415). However, lists compiled in this manner would also have to be applied to normal-hearing and hearing-impaired individuals to verify their equivalence.

In the current findings, the slope lists showed the greatest equivalence in listeners with normal peripheral hearing. This could be of great value in

assessing individuals whose primary complaint is an inability to understand speech in the presence of noise, despite normal peripheral hearing (Bellis, 2003b:10). In addition, the elimination of 4 lists from the collection improved the equivalence of these lists in the presence of a simulated high-frequency hearing loss, and a collection of 18 lists therefore remain that could be used to assess individuals with such a hearing loss. It is important, however, that normative data be established first and that the material is also validated on individuals with a hearing loss before any definite conclusions or diagnoses are made based on these results. Before evaluation with other filters or validation on individuals with different types of hearing loss (such as a low- or mid-frequency hearing loss), the results obtained with these lists when testing individuals with hearing loss should be interpreted with caution.

5.5 Conclusion

A final comparison of the results reported in this study to previously reported findings is shown in Table 5.18. The minimum, maximum and mean values reported in the studies listed are compared to the findings for the slope lists and PB lists as calculated after excluding lists that differed significantly from other lists in the collection. These findings reflect the data from the first experiment or unfiltered condition as discussed earlier, since previous studies do not report on an evaluation of their material in the presence of a simulated hearing loss. The comparison of findings in the table indicate that both the slope lists and PB lists showed a high degree of equivalence when compared to previous studies after excluding lists that differed significantly from others in the collection. The slope lists showed greater equivalence than the PB lists, especially with regards to the deviations from the overall mean. This conclusion should, however, be applied with caution and with careful consideration to the findings of the filtered condition, as the presence of a hearing loss could significantly affect the inter-list equivalence. Therefore, the value of validating test materials on listeners with a simulated or clinical hearing loss cannot be overstated.

Table 5.18: Comparison of slope lists and PB lists (excluding problematic lists) to existing literature

AUTHORS	STD DEV FROM MEAN (dB)	WITHIN-SUBJECT STD DEV (dB)	DEV FROM OVERALL MEAN (dB)
Plomp & Mimpfen (1979)	x	0.90	x
Nilsson et al (1994)	0.78	1.13	All within +/- 1dB
Kollmeier & Wesselkamp (1997)	0.27	x	x
Versfeld et al (2000)	0.56	1.07	x
Vaillancourt et al (2005)	0.50	1.10	All within +/- 1dB
Wong & Soli (2005)	1.00	1.80	All within +/- 1dB
Hällgren et al (2006)	1.10	x	All within +/- 1dB
Van Wieringen & Wouters (2006)	1.20	1.17	All within +/- 1dB
Wong et al (2007) MHINT-M	0.62	0.89	x
Wong et al (2007) MHINT-T	0.94	0.75	x
Minimum	0.27	0.75	
Maximum	1.20	1.80	All within +/- 1dB
Mean	0.77	1.10	
Slope lists (excluding lists 3,12,19,20)	0.46	1.14	All within +/- 1dB
PB lists (excluding lists 8,11,15,21)	0.64	1.22	All within +/- 1.5dB

The main aim of the present research was to develop a valid and reliable Afrikaans test of sentence recognition thresholds in noise. There are several different types of validity and reliability that could serve to describe the test performance of a developed measure. These are illustrated in the table below, along with the manner in which these issues were addressed in the current study.

Table 5.19: Different types of validity and reliability and application thereof to the current test

	TEST PERFORMANCE VARIABLE	APPLICATION IN PRESENT STUDY
VALIDITY	Content validity: <i>Extent to which test items reflect behaviour of interest</i>	Sentence material represents everyday speech better than single words Sentences rated for naturalness by native speakers Presence of noise typical of everyday listening situations
	Criterion-related validity: <i>Correlation between test score and other measures of the same behaviour</i>	No “gold standard” available in Afrikaans for comparison
	Construct validity: <i>Correlation with underlying theoretical constructs</i>	Developed test provides means of quantifying the ability to understand speech in the presence of noise, as this cannot be determined from pure tone audiogram
RELIABILITY	Coefficient of stability: <i>Test-retest reliability</i>	Re-test subjects after period of time – not completed during current project, but recommended for future project
	Coefficient of internal consistency: <i>First half of list yielding similar results to second half of list</i>	Analysis of variance conducted to determine effect of list on threshold value. Effect of list in final collections according to unfiltered results more significant for PB lists than slope lists (as indicated by smaller p-value)
	Coefficient of equivalence: <i>Same results across:</i> <ul style="list-style-type: none"> • <i>Different testers</i> • <i>Different lists</i> 	Pre-recorded material used to ensure equivalence across testers; inter-tester reliability not yet assessed Inter-list equivalence comparable to previously reported findings as demonstrated by results of Experiment I and II in the third phase

5.6 Summary

The current chapter provided a critical discussion of the results of the research in light of existing literature. The results of each of the three project phases were discussed separately. Implications for future research were indicated throughout, and limitations of current and previous studies identified.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

The main aim of this study was to develop a valid and reliable Afrikaans test of sentence recognition thresholds in noise. The data collection and analyses occurred in three phases, and the results attained during those phases are all presented and discussed in this report. The current chapter serves as the closing of the report by drawing conclusions from the reported results and critically reviewing the research process. Recommendations for further research are presented along with implications of the research.

6.2 Conclusions

The research process described in this report was primarily aimed at developing a valid and reliable Afrikaans test of sentence recognition thresholds in noise. To achieve this main aim, three distinct sub-aims were formulated to guide the research process. Each sub-aim was attained in a separate phase of the project. In addition, two specific research questions were raised in the first chapter. These two questions were indirectly answered throughout the course of the research project. The following sections elucidate the conclusions pertaining to each of the project aims (main aim and sub-aims) followed by conclusions related to the research questions.

6.2.1 Main aim: Development of a valid and reliable Afrikaans test for sentence recognition thresholds in noise

- ❑ The research process was successful in attaining the main aim set for this project. The final product was two sets of 18 sentence lists each with an added speech-weighted noise that can be used to determine the sentence recognition threshold in noise using a quick and simple procedure.

- ❑ The set of lists that were compiled according to the intelligibility of the sentences in noise (slope lists) showed the greatest inter-list reliability in normal-hearing subjects. In a group of subjects with simulated hearing loss, the phonetically balanced (PB) lists were slightly more equivalent. The slope lists remained the best option for assessing individuals with normal peripheral hearing, while the PB lists are the more prudent choice for hearing-impaired patients, as its spectral content is better controlled. Further research is necessary to assess the inter-list equivalence of the slope lists in patients with hearing impairments of a different configuration to the high frequency loss simulated in this study before using these lists for clinical evaluations of hearing-impaired individuals.
- ❑ The validity of the test lies in the fact that the sentences are natural and representative of everyday speech, as rated by native speakers of Afrikaans. The test material therefore shows good face validity in measuring an individual's ability to cope with typical everyday speech. In addition, the presence of background noise ensures that the test is conducted in the type of listening situation that commonly occurs in daily life and constitutes a major challenge for individuals with hearing impairment, thereby providing a valid measure of everyday auditory functioning.
- ❑ The reliability of the test was verified by applying the sentence collection not only to group of normal-hearing young adults, but also to listeners with a simulated hearing loss. The results showed a high degree of inter-list equivalence (reliability), whether the lists were arranged according to phonetic content or purely according to intelligibility in noise. A number of lists were excluded from each set to further improve this uniformity.

6.2.2 Sub-aim 1 (Phase I): Development of a collection of recorded Afrikaans sentences suitable for the assessment of speech recognition in noise

- ❑ The combination of translated and original material proved successful in compiling a sentence collection of equivalent grammatical complexity and naturalness.
- ❑ The characteristics specified for the sentence content (complete sentences, representative of everyday speech, and free from proverbs, questions, exclamations and proper nouns) ensured that the material received a high rating of naturalness and very few changes or exclusions (13 in total) were necessary to improve this aspect.
- ❑ The rating of grammatical complexity and naturalness showed the sentence collection to be uniform in terms of complexity and representative of everyday speech.

6.2.3 Sub-aim 2 (Phase II): Selecting from the recorded material a collection of sentences with equivalent intelligibility in the presence of noise

- ❑ The method of eliminating sentences instead of re-scaling intensities proved reliable and efficient. A smaller sample size and less experimentation time were required as it was not necessary to re-test the same materials repeatedly. The size of the sentence collection was reduced by 57% (slightly more than the initially predicted 50%), which corresponded well with previous studies. The final collection of sentences also proved to yield an SNR at 50% recognition with a mean value and standard deviation that agreed with previous reports, indicating that this method is appropriate and reliable.
- ❑ A significant effect of gender on test results was found during the first equalisation procedure, but this did not occur during the second procedure. The small sample size and the fact that this effect has not been previously documented indicate that these results cannot be generalised yet. However, previous studies did not control for or

investigate this effect and there is some evidence in the literature of gender differences in terms of noise perception and auditory working memory (Ruytjens et al., 2007:2074). Therefore, this effect requires further investigation with a larger sample of subjects, and future researchers developing tests of speech recognition in noise should control for gender effects when selecting research subjects.

- ❑ Within the present collection, the grammatical complexity of the sentences did not have a significant effect on intelligibility. This may be due in part to the uniformity of the collection in terms of grammatical level.
- ❑ Practice did not have a statistically significant effect on test performance, but the use of practice lists was nevertheless recommended in the light of previously reported findings.

6.2.4 Sub-aim 3 (Phase III): Comparing inter-list reliability and response variability of two list sets compiled using two different methods of list compilation

- ❑ The method developed to compile phonetically balanced lists proved to be highly effective when compared to techniques reported in previous studies. Compared to the phonetically balanced lists of the current study, the lists compiled purely according to their intelligibility slopes (slope lists) revealed a much smaller percentage of phoneme counts that were within +/-1 phoneme from the ideal count. This percentage nevertheless corresponded well with previous studies that compiled lists based on phonetic content. The current method of phonetic balance therefore showed itself to be reliable, but the phonetically balanced lists of some previous studies did not show a greater degree of phonetic balance than that which occurred incidentally when arranging lists according to their intelligibility in noise.
- ❑ The experimental method where lists were arranged according to the intelligibility slopes of the sentences was found to be very time efficient, as only 14 exchanges were needed to obtain a high degree of equivalence between lists (maximum difference of 7% between best and

worst scored list at each of the three SNRs used during the second phase). The PB lists were less equivalent in this regard (differences between best and worst scored lists ranged from 10-17% at different SNRs).

- ❑ Within a normal-hearing research sample, both sets of lists showed inter-list equivalences that corresponded well with previously reported findings in terms of standard deviation of the lists, within-subject deviation as well as deviations from the overall mean. The slope lists showed greater equivalence than the PB lists in normal-hearing listeners (Experiment I).
- ❑ In the presence of a simulated high frequency hearing loss (Experiment II), the PB lists showed a slightly greater degree of inter-list equivalence than the slope lists. However, it was possible to use the findings of the second experiment to remove from both collections lists that differed significantly from other lists in the collection, thereby improving the uniformity of both sets. The same number of lists ($n = 4$) had to be removed from both sets to eliminate significant inter-list differences, thereby producing the same number of lists for both sets (18 lists each).
- ❑ The findings of the second experiment demonstrated that lists found to be of equivalent difficulty in normal-hearing listeners might show significant differences in the presence of a hearing loss. This was especially true for lists that are not phonetically balanced. However, within the set of PB lists, there were also differences in list equivalence between the unfiltered (normal-hearing listeners) and filtered (simulated hearing loss) results. Therefore, phonetic content alone does not fully account for list equivalence.
- ❑ After removing from the two sets all the lists that showed significant inter-list differences, both collections showed an adequate degree of inter-list reliability when compared to existing literature, with the slope lists showing greater equivalence in normal-hearing listeners and the PB lists performing slightly better with the simulated loss.

- ❑ The findings indicate the importance of verifying inter-list reliability not only in normal-hearing listeners, but also in listeners with a simulated or clinical hearing loss.

6.2.5 Research question 1: What methods for the development of a test for sentence recognition in noise have been documented in the literature, and how successful were these methods?

- ❑ The different methods that have been used by previous researchers in the development of a test for speech recognition in noise were extensively explored and critically discussed in Chapter 2 to provide a framework that served to guide the methodology of the present study (see Figure 2.1).
- ❑ The validity, reliability, and efficiency of the current and previously reported methods were evaluated through the experimentation conducted in the different phases of the project. The findings from each phase were discussed in this chapter under the three different sub-aims (6.2.2, 6.2.3, and 6.2.4).

6.2.6 Research question 2: Is it possible to improve or streamline previously reported methods that will make the development of such a test more efficient while still producing a reliable measure?

- ❑ During the second phase of the project, sentences that deviated considerably from the mean intelligibility of all the sentences at a particular SNR were eliminated from the collection. This is in contrast to the commonly used method of re-scaling the intensities of such sentences and re-testing their intelligibility with another group of subjects (Plomp and Mimpen, 1979:44; Nilsson et al., 1994:1088; Kollmeier and Wesselkamp, 1997:2414; Wong and Soli, 2005:279; Hällgren et al., 2006:229; Wong et al., 2007:71S). The current method proved to yield a highly equivalent collection of sentences, large enough to compile 22 lists of 10 sentences each, without the need for multiple rounds of testing that would have required a greater number of research subjects and more experimentation. In this regard, the present study was successful in

streamlining a specific aspect of the methodology followed in the development of the test.

- ❑ The commonly used method of compiling phonetically balanced lists and assessing these lists only in a normal-hearing sample was demonstrated to be a weakness in many previous reports. The use of intelligibility slopes to compile equivalent sentence lists was shown to be both simple and reliable, especially when used in listeners with normal peripheral hearing. In addition, it was found possible to improve the inter-list reliability of these lists according to findings made when testing listeners with a simulated hearing loss.

6.3 Implications of findings

The results of the current research have three main implications. First of all, the output of the project is a collection of sentence lists that are of equivalent difficulty and that can be used to evaluate the ability of Afrikaans listeners to understand sentences in the presence of noise. In the South African context there is an extreme dearth of pre-recorded materials available for the evaluation of speech perception, especially in languages other than English. The test developed during this project therefore constitutes a valuable resource to audiologists in South Africa, as it provides clinicians with the means to evaluate an essential communication skill. Not only can audiologists use this measure to quantify a common problem in patients with hearing impairment, but it is also possible to repeatedly assess this area using the different sentence lists in order to monitor progress or evaluate adjustments made to amplification devices.

Secondly, the methodology followed during the development of this collection could serve as a template for the development of similar tests in other languages. The exposition and critical discussion of methodologies reported in the literature offer important guidelines for future developers of similar tests. The application and evaluation of these methods provide further insights on the reliability of the proposed methodology. The selection criteria specified for

the listeners and speakers, the scoring method (syllable scoring) used in initial phases, as well as the novel method of list compilation according to intelligibility slopes are all unique features of the developed methodology that could be of value to future researchers. In addition, the procedures followed to compile the PB lists have been reported in great detail and were shown to be very accurate, and can therefore guide future developers of similar tests in compilation of such lists.

Thirdly, the comparison of two list compilation methods yielded valuable information about the reliability of these methods. These findings indicated that lists that are compiled purely according to phonetic content are not necessarily of equivalent intelligibility in noise. Furthermore, it was demonstrated that sentence lists that exhibit equivalent intelligibility in normal hearers, are not necessarily equivalent for listeners with a hearing loss. This demonstration underscores the importance of assessing list equivalence or inter-list reliability in a population with a hearing loss (simulated or clinical).

6.4 Critical evaluation of research

The procedures followed during the research project are critically evaluated in Table 6.1. The table indicates the different options for each variable as reported in existing literature. The options selected for the study at hand are shaded. The strengths and limitations of the selected methods are discussed in the last two columns.

Table 6.1: Critical evaluation of test method variables as applied in current study

TEST VARIABLE	OPTIONS FOR EACH VARIABLE (REPORTED IN LITERATURE)	STRENGTHS	LIMITATIONS	
STIMULUS	Composition of speech material	Develop own/original material Adaptation of existing material	Combination of methods was successful in producing a large collection of sentences with great uniformity in terms of grammatical complexity. Sentence lengths showed greater variation than many other studies.	
	Equalising sentence difficulty	Re-scale intensity of sentences that are too hard / too easy Eliminating / excluding sentences that are too hard / too easy	Required smaller amount of subjects and time without reducing size of collection beyond previously reported reductions.	Required three rounds of testing, each time with different subjects.
		Select subset or decide on re-scaling based on SNR-50 only Select subset or decide on re-scaling based on SNR-50 and psychometric slope	Takes into account the slope of the psychometric function and thereby gives a more accurate reflection of changes in threshold.	
		Noise type	Recorded speech (continuous discourse or multi-talker babble) Idealised speech weighted noise consistent with the mean LTASS spectrum across languages Noise created according to LTASS of recording specific to test being developed	
	Gender of speaker	Male Female Male and female speaker	Only one set of recordings needed, therefore less testing time as subjects did not have to be evaluated with two sets of sentence material.	Would be useful to have the final set of sentences recorded using a male voice as well. Results could then be generalised to a greater number of everyday situations.
	Training of speaker	Speech therapist / audiologist	Speaker had knowledge and experience of correct articulation and of speech audiometry.	

Table 6.1: Critical evaluation of test method variables as applied in current study (continued)

PRESENTATION	Presentation method	Fixed presentation level in initial phases Fixed presentation level throughout Adaptive presentation method once lists have been compiled Adaptive presentation method throughout	Initial fixed level simplified testing during second phase, whilst adaptive testing in the third round served to concentrate presentation levels in the range that yields the smallest standard deviations in SNR-50.	
	Presentation level (speech and noise pre-mixed and re-scaled)	70 dB SPL	Presentation level of the noise is a non-critical factor in speech tests and can be chosen arbitrarily, as long as the noise presentation level exceeds the individual's threshold (Wagener, 2004).	
	Auditory transmission channel	Sound-field Headphones (binaural) Headphones (monaural) Headphones simulating noise front and side conditions	Headphone testing more reliable and less variable than sound-field testing, and no need for site-specific norms (Vaillancourt et al., 2005).	Headphone norms cannot be applied to results obtained with sound-field testing, which will be used on hearing aid or cochlear implant users. Sound-field norms will need to be established prior to testing such individuals.
SUBJECT	Subject characteristics	Auditory characteristics: Thresholds ≤ 15 dB HL from 250 to 8000 Hz, normal otoscopic examination, tympanograms, otologic history, no history/symptoms of auditory processing disorder Age range: 18-30 years Language: Native speaker, Grade 12 education in test language Cognition: Completed Grade 12 in mainstream education	Selection criteria ensured homogeneity of research sample and controlled factors that could possibly affect ability to understand speech in noise.	Required a series of selection procedures.
RESPONSE	Response channel	Written / typed Verbal	Reduced testing time. Errors could be marked on test form, and this provided a text version of responses for further analyses.	
	Scoring method initial phase	Word-by-word Syllable-by-syllable	Gives more detailed account of performance on each sentence. Especially valuable due to conjunctive spelling tendencies of Afrikaans.	Not previously documented.

Table 6.1: Critical evaluation of test method variables as applied in current study (continued)

TEST PERFORMANCE	VALIDITY	Face validity	High degree of apparent validity for estimating an individual's ability to deal with typical speech stimuli	Sentences as stimuli have a high degree of face validity. Presence of background noise also representative of typical everyday listening environments.	Only one speaker – limited generalisation to all listening situations. Type of noise used not typical to everyday listening.
		Content validity	Test material considered natural by native speakers of the test language	Sentences judged to be natural by native speakers from different backgrounds.	
			Use noise as part of the test material to make test condition more representative of everyday listening environment	Presence of noise improved content validity.	
		Criterion-related validity	Compare test results with existing tests	Developed measure makes a valuable contribution to the field as a useful tool to evaluate speech recognition in noise.	No standardised tests of speech recognition in noise available in Afrikaans.
			No existing tests to compare with - refer to construct validity		
	Construct validity	Apply test to hearing-impaired population with deficit in evaluated skill Simulate hearing loss in normal-hearing population	Current study assessed test material using both normal-hearing listeners and listeners with a simulated hearing loss, yielding insightful findings on validity and reliability.	Test material only applied to listeners with a simulated high frequency hearing loss. More experimentation needed to validate material for low or mid-frequency losses. Ideally, material should also be validated in hearing-impaired populations.	
	RELIABILITY	Stability	Test-retest reliability evaluated by re-testing subjects after a period of time	Subjects in third phase were re-tested, but during second experiment a hearing loss was simulated.	Test-retest reliability not assessed as part of current project – to be conducted as a separate study.
		Equivalence / Inter-list reliability	Achieved by phonetically balancing lists Achieved by arranging lists according to intelligibility slopes Assess by comparing mean score for each list across subjects with overall mean Assess according to standard deviation of list means across subjects	Both methods of list compilation applied and evaluated on normal-hearing subjects as well as listeners with simulated hearing loss. Inter-list reliability / equivalence extensively explored and improved through removal of lists that differed significantly from others in the collection.	Inter-list reliability in normal listeners better for slope lists, but in listeners with simulated loss phonetically balanced lists showed greater equivalence.
			Internal consistency	Compare thresholds obtained by a certain combination of lists with thresholds from a different combination of lists Compare thresholds obtained by one list to those obtained with other lists by an analysis of variance	ANOVA procedures supplied the necessary insights to enable researcher to eliminate from each collection those lists that differed significantly from other lists.

SENSITIVITY	Sensitivity / specificity	<p>Determine statistical power to predict a true difference in threshold</p> <p>Apply test to both normal-hearing and hearing-impaired individuals and assess ability of test to separate these groups</p>	<p>Lists applied to normal listeners and same listeners with simulated hearing loss. Significance of difference between results for these two conditions compared and found to be significant for all lists, indicating that the material was sensitive to changes in threshold, albeit simulated.</p>	Sensitivity / specificity not assessed in a clinically impaired population.
-------------	---------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------

6.5 Recommendations for further research

- ❑ The experimentation conducted in this project confirmed the content and construct validity as well as the inter-list reliability of the developed test. However, additional research could serve to further refine and standardise the test. The developed measure could also stimulate further research as it now provides a method of assessing sentence recognition in noise for Afrikaans-speaking listeners, which could be compared to existing tests in the language.
- ❑ The size of the developed sentence collection may be increased by re-scaling some of the sentences rejected during the second phase and then re-evaluating their equivalence through experimental application.
- ❑ The test-retest reliability of both sets of lists should be assessed. This would indicate the coefficient of stability of the test (Ostergard, 1983:224). With sentence material, it can be expected that the listener may become familiar with the material due to its redundancy and therefore perform better during the second test (Nilsson et al., 1994:1085). Allowing a period of time (a month or two, as demonstrated by Cameron and Dillon, 2007b) between testing and re-testing could be a way of reducing this practice effect.
- ❑ The possibility of compiling lists that are balanced in terms of frequency spectrum could be explored and compared to the use of phonetic content as an indication of frequency content. This could be achieved by determining the spectral content of each list and exchanging lists on a trial-and-error basis in order to balance the spectral content of the lists.
- ❑ Site-specific norms should be established for free-field testing using the developed measure. The influence of room acoustics could cause variability in the signal, and it is therefore necessary to establish norms specific to a particular acoustic environment when testing in the sound-field (Vaillancourt et al., 2005:365). This would enable researchers to use the developed test to assess individuals with hearing aids or cochlear implants.

- ❑ Gender differences in speech recognition in noise could be further investigated. The findings of the current research indicated the possibility of a gender difference in sentence recognition in noise. However, the results were not conclusive and this effect is not reported elsewhere, although there have been reports of gender differences in certain auditory processing skills (Ruytjens et al., 2007:2074). Further research using a larger research sample is needed to clarify this issue.
- ❑ The inter-list reliability of both sets when applied to individuals with varying audiogram configurations and/or modes of application requires investigation. The experiments reported in the current study only evaluated inter-list equivalence in normal-hearing subjects and in the presence of a simulated high frequency hearing loss. The findings indicated that not all the lists were of equivalent intelligibility in the presence of this type of hearing loss. Adjustments were made to the collection of lists to compensate for this fact. However, hearing impairments affecting other frequency areas, as well as modes of amplification that do not provide equal audibility across the entire speech frequency range, may have a different effect on this equivalence. For this reason, more experimentation on hearing-impaired individuals, with and without amplification, is needed before clinical use of the test on this population.
- ❑ The establishment of norms for speech recognition thresholds using the developed sentences in quiet would enable an additional use of the material. Sentence materials are more representative of everyday speech than the single words often used to obtain speech recognition thresholds, and can therefore provide a more representative reflection of a patient's everyday listening skills.
- ❑ Sentence recognition abilities of Afrikaans listeners as assessed with the developed material could be compared to results attained with traditional measures of speech recognition in Afrikaans (spondaic thresholds and word discrimination lists). The use of a variety of materials (such as words and sentences) will allow clinicians to obtain a more holistic impression of a patient's ability to hear speech signals. With a more accurate and

comprehensive view of the patient's everyday functioning in terms of speech comprehension, both counselling and intervention could be refined to suit the specific needs of each individual.

- ❑ The developed material may be used to assess individuals with normal peripheral hearing complaining of a reduced ability to understand speech in noise to determine whether the test is able to validate and/or quantify these patients' complaints.
- ❑ The sensitivity and specificity of the developed measure should be assessed in cochlear implant users. The establishment of an Afrikaans test of speech recognition (both in quiet and in noise) could be of great value to clinicians working with Afrikaans-speaking patients using cochlear implants. It will permit audiologists with a means to assess important auditory skills in this population in their first language, both as a baseline measurement and as a quantification of progress. Changes made to the settings of the implant could also be assessed to determine its effect on the patient's speech recognition.

6.6 Final conclusion

Measuring a patient's ability to discern speech in noise not only quantifies one of the main complaints of individuals with hearing impairment, but also provides the audiologist with valuable information needed for successful rehabilitation of these patients (Smits et al., 2006:538). Therefore, the current study makes a valuable contribution within the South African context. Not only does it provide audiologists with the means to assess a critical communication function in Afrikaans-speaking patients, it also lays the groundwork for the development of similar tests in other South African languages. The expansion of the existing battery of speech audiometric tests and consequential enhancement of diagnostic procedures can improve service delivery to the hearing-impaired population of South Africa.

REFERENCES

- American Speech-Language-Hearing Association. (2005). *Guidelines for Manual Pure-Tone Threshold Audiometry* [Guidelines]. Retrieved September 19, 2007, from <http://www.asha.org/policy>
- Bamiou, D. (2007). Measures of Binaural Interaction. In F.E. Musiek and G.D. Chermak (Eds.), *Handbook of (Central) Auditory Processing Disorder: Auditory Neuroscience and Diagnosis Volume I* (pp. 257-286). San Diego: Plural Publishing Inc.
- Barab, S. & Squire, K. (2004). Design-Based Research: Putting a Stake in the Ground. *The Journal of the Learning Sciences*, 13(1), 1-14.
- Barfod, J. (1979). Speech Perception Processes and Fitting of Hearing Aids. *Audiology*, 18, 430-441.
- Barrenäs, M. & Wikström, I. (2000). The Influence of Hearing and Age on Speech Recognition Scores in Audiological Patients and in the General Population. *Ear and Hearing*, 21(6), 569-577.
- Bayles, K.A. & Kaszniak, A.W. (1987). *Communication and Cognition in Normal Aging and Dementia*. Boston: College-Hill Press.
- Bellis, T.J. (2003a). *Assessment and Management of Central Auditory Processing Disorders in the Educational Setting From Science to Practice* (2nd ed.). New York: Thomson Delmar Learning.
- Bellis, T.J. (2003b). Auditory processing disorders: It's not just kids who have them. *The Hearing Journal*, 56(5), 10-18.

- Bench, J. & Bamford, J. (1979). *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children*. London: Academic Press.
- Bess, F.H. (1983). Clinical Assessment of Speech Recognition. In D.F. Konkle & W.F. Rintelmann (Eds.), *Principles of Speech Audiometry* (pp. 127-202). Baltimore: University Park Press.
- Blandy, S. & Lutman, M. (2005). Hearing threshold levels and speech recognition in noise in 7-year-olds. *International Journal of Audiology*, 44(8), 435-443.
- Boersma, P. & Weenink, D. (2006). *Praat: doing phonetics by computer (Version 4.3.14)* [Computer program]. Retrieved June 20, 2006, from <http://www.praat.org/>
- Brand, T. & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, 111(6), 2801-2810.
- Buitendag, M.M. (1994). Die opstel en standardisering van 'n Afrikaanse Reseptiewe Woordeskattoets. Unpublished M Logopaedics dissertation. Department of Logopaedics, University of Pretoria.
- Burke, L.E. & Nerbonne, M.A. (1978). The influence of the guess factor on the speech reception threshold. *Journal of the American Auditory Society*, 4(3), 87-90.
- Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., et al. (1994). An international comparison of long-term average speech spectra. *Journal of the Acoustical Society of America*, 96(4), 2108-2120.
- Cameron, S. & Dillon, H. (2007a). Development of the Listening in Spatialized Noise-Sentences Test (LISN-S). *Ear and Hearing*, 28(2), 196-211.

- Cameron, S. & Dillon, H. (2007b). The listening in spatialized noise-sentences test (LISN-S): test-retest reliability study. *International Journal of Audiology*, 46(3), 145-153.
- Carhart, R. (1968). Future horizons in audiological diagnosis. *The Annals of Otolaryngology, Rhinology & Laryngology*, 77, 706-716.
- Carhart, R. (1970). Discussion, questions, answers, comments. In C. Røjskjer (ed.), *Speech Audiometry* (p. 229). Second Danavox Symposium, Andelsbogtrykkeriet i Odense, Denmark.
- Carstens, W.A.M. (2003). *Norme vir Afrikaans: Enkele riglyne by die gebruik van Afrikaans* (4th ed.). Pretoria: Van Schaik Uitgewers.
- Cloete, J. (1997). 'n Afrikaanse vertaling van die BKB-sinne. Unpublished B Speech-Language and Hearing Therapy dissertation. Department of Speech-Language and Hearing Therapy, University of Stellenbosch.
- Crandell, C.C. (1991). Individual Differences in Speech Recognition Ability: Implications for Hearing Aid Selection. *Ear and Hearing*, 12(6), Supplement, 100S-108S.
- Davis, H. (1970). Abnormal Hearing and Deafness. In H. Davis & S.R. Silverman (Eds.), *Hearing and Deafness* (3rd ed.) (pp. 83-139). New York: Holt, Rinehart, and Winston, Inc.
- Department of Arts and Culture. (2002). *National Language Policy Framework*. Retrieved January 5, 2007, from <http://www.info.gov.za/otherdocs/2002/langpolframe.pdf>
- De Villiers, M. & Ponelis, F.A. (1987). *Afrikaanse klankleer* (revised ed.). Cape Town: Tafelberg.

- DiStefano, P. & Valencia, C. (1980). The Effects of Syntactic Maturity on Comprehension of Graded Reading Passages. *The Journal of Educational Research*, 73(5), 247-251.
- Feuerstein, J.F. (1992). Monaural versus Binaural Hearing: Ease of Listening, Word Recognition, and Attentional Effort. *Ear and Hearing*, 13(2), 80-86.
- Gatehouse, S. & Robinson, K. (1997). Speech tests as measure of auditory processing. In M. Martin (ed.), *Speech Audiometry* (2nd ed.) (pp. 74-88). London: Whurr Publishers Ltd.
- Geffner, D. (2007). Central Auditory Processing Disorders: Definition, Description, and Behaviors. In D. Geffner and D. Ross-Swain (Eds.), *Auditory Processing Disorders: Assessment, Management, and Treatment*. (pp.25-48). San Diego: Plural Publishing Inc.
- Gelfand, S. A. (2001). *Essentials of Audiology* (2nd ed.). New York: Thieme.
- Hällgren, M., Larsby, B. & Arlinger, S. (2006). A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *International Journal of Audiology*, 45(4), 227-237.
- Hammond, S.S.J. (1987). An evaluation of a signal to noise ratio procedure for speech discrimination testing. Unpublished BSc Logopaedics dissertation. Department of Logopaedics, University of Cape Town.
- Hannley, M. (1986). *Basic Principles of Auditory Assessment*. London: Taylor and Francis Ltd.
- Hegde, M.N. (2003). *Clinical Research in Communicative Disorders: Principles and Strategies* (3rd ed). Texas: Pro-ed.

- Hernvig, L.H. & Olsen, S. Ø. (2005). Learning effect when using the Danish Hagerman sentences (Dantale II) to determine speech reception threshold. *International Journal of Audiology*, 44(9), 509-512.
- Huysamen, G.K. (1994). *Methodology for the Social and Behavioural Sciences*. Johannesburg: International Thomson Publishing.
- International Organization for Standardization. (1991). *Specification for standard reference zero for the calibration of pure-tone air conduction audiometers*. ISO 389. Geneva: ISO.
- Kalikow, D.N., Stevens, K.N. & Elliott, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61, 1337-1351.
- Killion, M.C. (2002). New thinking on Hearing in Noise: A Generalized Articulation Index. *Seminars in Hearing*, 23(1), 57-75.
- Killion, M.C. & Niquette, P.A. (2000). What can the pure-tone audiogram tell us about a patient's SNR loss? *The Hearing Journal*, 53(3), 46-53.
- Kollmeier, B. & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *Journal of the Acoustical Society of America*, 102(4), 2412-2421.
- Konkle, D.F. & Rintelmann, W.F. (1983). *Principles of Speech Audiometry*. Baltimore: University Park Press.
- Le Roux, T.H. & Pienaar, P. de V. (1976). *Uitspraakwoordeboek van Afrikaans*. Pretoria: J.L. van Schaik.
- Leedy, P.D. & Ormrod, J.E. (2005). *Practical Research: Planning and Design* (8th ed.). New Jersey: Pearson Education Inc.

- Lucks Mendel, L. & Danhauer, J.L. (1997). *Audiologic Evaluation and Management and Speech Perception Assessment*. San Diego: Singular Publishing Group, Inc.
- Lutman, M.E. (1997). Speech tests in quiet and noise as a measure of auditory processing. In M. Martin (ed.), *Speech Audiometry* (2nd ed.) (pp. 63-73). London: Whurr Publishers Ltd.
- Lutman, M.E. & Clark, J. (1986). Speech identification under simulated hearing aid frequency response characteristics in relation to sensitivity, frequency resolution, and temporal resolution. *Journal of the Acoustical Society of America*, 80(4), 1030-1040.
- Lyregaard, P. (1997). Towards a theory of speech audiometry tests. In M. Martin (Ed.), *Speech Audiometry* (2nd ed.) (pp. 34-62). London: Whurr Publishers Ltd.
- Mackersie, C.L. (2002). Tests of speech perception abilities. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 10, 392-397.
- Martin, F.N., Champlin, C.A. & Perez, D.D. (2000). The question of phonetic balance in word recognition testing. *Journal of the American Academy of Audiology*, 11(9), 489-493.
- Maxwell, D.L. & Satake, E. (2006). *Research and Statistical Methods in Communication Sciences and Disorders*. Canada: Thomson Delmar Learning.
- McLauchlin, R.M. (1980). Speech Protocols for Assessment of Persons With Limited Language Abilities. In R.R. Rupp & K.G. Stockdell, Sr (Eds.), *Speech Protocols in Audiology* (pp. 253-286). New York: Grune & Stratton Inc.

- Meister, H., Von Wedel, H. & Walger, M. (2004). Psychometric evaluation of children with suspected auditory processing disorders (APDs) using a parent-answered survey. *International Journal of Audiology*, 43(8), 431-437.
- National Council on the Aging. (1999). *The consequences of untreated hearing loss in older persons*. Retrieved January 4, 2007, from <http://www.ncoa.org/attachments/UntreatedHearingLossReport%2Epdf>
- National Institute on Deafness and Other Communication Disorders. (2007). *Statistics about Hearing Disorders, Ear Infections, and Deafness*. Retrieved February 28, 2007, from <http://www.nidcd.nih.gov/health/statistics/hearing.asp>
- Naudé, E. (1998). Gegewens oor taalontwikkeling by Afrikaanssprekende kinders. *Clinica: Applications in Clinical Practice of Communication Pathology*, 3, 73-105.
- Neijenhuis, K.A.M., Stollman, M.H.P., Snik, A.F.M. & Van den Broek, P. (2001). Development of a Central Auditory Test Battery for Adults. *Audiology*, 40(2), 69-77.
- Nilsson, M.J., 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.
- Northern, J.L. and Downs, M.P. (2002). *Hearing in Children* (5th ed.). Philadelphia: Lippincott Williams & Wilkins.
- Olivier, J.M. (2000). Die vertaling en ontwikkeling van sinsmateriaal vir die evaluasie van spraakpersepsie by Xhosa-sprekendes. Unpublished M Communication Pathology dissertation. Department of Communication Pathology, University of Pretoria.

- Ostergard, C.A. (1983). Factors influencing validity and reliability of speech audiometry. *Seminars in Hearing*, 4(3), 221-240.
- Owens, E. (1983). Speech Recognition and Aural Rehabilitation. In D.F. Konkle & W.F. Rintelmann (Eds.), *Principles of Speech Audiometry* (pp 353-374). Baltimore: University Park Press.
- Pakendorf, C. (1998). 10-punt plan vir die vertaling en kulturele aanpassing van toetsmateriaal binne die Suid-Afrikaanse konteks. *Clinica: Applications in Clinical Practice of Communication Pathology*, 1998, 1-9.
- Persson, P., Harder, H., Arlinger, S. & Magnuson, B. (2001). Speech Recognition in Background Noise: Monaural versus Binaural Listening Conditions in Normal-hearing Patients. *Otology and Neurotology*, 22(5), 625-630.
- Plomp, R. & Mimpen, A.M. (1979). Improving the Reliability of Testing the Speech Reception Threshold for Sentences. *Audiology*, 18, 43-52.
- Polit, D.F. & Hungler, B.P. (1991). *Nursing Research: Principles and methods*. Philadelphia: J.B. Lipincott.
- Roets, R. (2005). *Spraakoudiometrie in Suid-Afrika : ideale kriteria teenoor kliniese praktyk*. Unpublished M Communication Pathology dissertation. Department of Communication Pathology, University of Pretoria.
- Roush, J. (2001). *Screening for hearing loss and otitis media in children*. San Diego: Singular-Thomson Publishing Group.
- Rupp, R.R. (1980). Determination of the Spondee Threshold: Classical Approaches. In R.R. Rupp & K.G. Stockdell, Sr (Eds.), *Speech Protocols in Audiology* (pp. 67-97). New York: Grune & Stratton Inc.

- Rupp, R.R. & Stockdell, K.G., Sr. (1980). The Roles of Speech Protocols in Audiology. In R.R. Rupp & K.G. Stockdell, Sr (Eds.), *Speech Protocols in Audiology* (pp. 5-39). New York: Grune & Stratton Inc.
- Ruytjens, L., Georgiadis, J.R., Holstege, G., Wit, H.P., Albers, F.W.J., & Willemsen, A.T.M. (2007). Functional sex differences in human primary auditory cortex. *European Journal of Nuclear Medicine and Molecular Imaging*, 34(2), 2073-2081.
- Salkind, N.J. (2006). *Exploring Research* (6th ed). New Jersey: Pearson Education Inc.
- Scott, T., Green, W.B., & Stuart, A. (2001). Interactive Effects of Low-Pass Filtering and Masking Noise on Word Recognition. *Journal of the American Academy of Audiology*, 12(9), 437-444.
- Silverman, S.R. (1983). Historical Foundations of Speech Audiometry. In D.F. Konkle & W.F. Rintelmann (Eds.), *Principles of Speech Audiometry* (pp 11-24). Baltimore: University Park Press.
- Silverman, S.R. & Hirsh, I.J. (1955). Problems related to the use of speech in clinical audiometry. *Annals of Otology, Rhinology, and Laryngology*, 64(4), 1234-1244.
- Smits, C., Kramer, S.E., & Houtgast, T. (2006). Speech Reception Thresholds in Noise and Self-Reported Hearing Disability in a General Adult Population. *Ear and Hearing*, 27(5), 538-549.
- Speaks, C. & Jerger, J. (1965). Method for measurement of speech identification. *Journal of Speech and Hearing Research*, 8, 185-194.
- Statistics South Africa. (2001). *Census 2001 Key Results*. Retrieved September 8, 2006 from <http://www.statssa.gov.za/>

APPENDIX A: INFORMED CONSENT FORM

Indien u gewillig is om aan die navorsingstudie deel te neem, teken asseblief die volgende ooreenkoms:

Ek, _____, verklaar hiermee dat ek toestem tot deelname aan die studie hierbo beskryf. Ek verstaan wat van my verwag sal word tydens die studie en dat deelname volkome vrywillig en konfidensiëel is. Ek verstaan ook dat ek die reg het om op enige tydstip my deelname aan die studie te onttrek.

Geteken: _____

Plek: _____ Datum: _____

APPENDIX C: PROFILE OF PARTICIPANT (GROUP A)

Date of Birth: _____

Highest academic qualification: _____

Current occupation: _____

Town/province of current address: _____

List towns and provinces where you attended school:

	Town	Province	Number of years
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

APPENDIX D: RATING OF SPEAKERS

Spreker no. _____

Beoordeelaar no. _____

Verstaanbaarheid	Goed	Gemiddeld	Swak	
Natuurlikheid	Goed	Gemiddeld	Swak	
Artikulasie	Goed	Gemiddeld	Swak	
Stemkwaliteit	Goed	Gemiddeld	Swak	
Resonansie	Goed	Gemiddeld	Swak	
Intonasie	Goed	Gemiddeld	Swak	
Spraakspoed	Te vinnig	Gepas	Te stadig	
Geaffekteerd/opgesmuk	Ja	Nee		
Dialekties	Ja	Nee		
Algehele indruk	/10			

APPENDIX E: CASE HISTORY FORM (GROUPS D-G)

Subject number: _____

1. Age: _____
2. Language spoken most often at home: _____
3. Grade 12 in Afrikaans, mainstream school: Yes No
4. History of:
 - a. Recurrent otitis media: Yes No
 - b. Excessive noise exposure Yes No
 - c. Neurologic disease (tumor/stroke): Yes No
 - d. Neurosurgery: Yes No
 - e. Traumatic Brain Injury: Yes No
 - f. Childhood APD / learning problems Yes No
5. Difficulty hearing speech in noise or fast speech that is so significant that you have consulted or considered to consult a professional about this?
 Yes No

Normal otoscopic examination: Yes No

Normal tympanometry: Yes No

Hearing thresholds ≤ 15 dB HL:

250 Hz: Yes No

500 Hz: Yes No

1000 Hz: Yes No

2000 Hz: Yes No

4000 Hz: Yes No

8000 Hz: Yes No

APPENDIX F: INSTRUCTIONS FOR RATING OF NATURALNESS

Geagte Deelnemer

Aangeheg is 'n lys Afrikaanse sinne wat ontwikkel is as deel van 'n toets vir spraakbegrip in die teenwoordigheid van agtergrondsgeraas. Sommige van die sinne is uit Engels vertaal, en die res is saamgestel vanuit Afrikaanse woordelyste.

U as deelnemer word vriendelik versoek om elkeen van die sinne op 'n skaal van een tot sewe te gradeer in terme van die *natuurlikheid* daarvan. Op die skaal verteenwoordig "1" dat die sin baie kunsmatig voorkom (met ander woorde dit is 'n swak gradering) en "7" dat die sin soos 'n natuurlike (met ander woorde realistiese of normale) Afrikaanse sin klink.

Indien u 'n gradering laer as 6 vir 'n bepaalde sin gee, word u versoek om 'n voorstel vir 'n meer realistiese of natuurlike sin in die regterkantste kolom te verskaf.

Voorbeeld:

#	Sin	1tot7	Voorstel vir verbetering
1	Die hond sit op die mat.	7	
2	Die pa lees sy koerant.	4	My pa lees sy koerant.

Byvoorbaat dankie vir u kosbare bydrae!

APPENDIX G: TEST FORM - SLOPE LISTS

	Lys 1			Lys 2	
1	Die hanswors het 'n snaakse gesig.		11	Die seun het die speletjie geken.	
2	Sy sny met 'n mes.		12	Kersfees is in die somer.	
3	Die huis het nege kamers.		13	Die polisie het die kar gejaag.	
4	My pa sluit die voorhek.		14	Die muis hardloop na sy gat toe.	
5	Hulle kyk na die horlosie.		15	Die vrou maak 'n speelding.	
6	Die sak sleep op die grond.		16	Daar is 'n hoop hout onder die boom.	
7	Die seun loop op sy hande.		19	Hulle sê 'n klomp lawwe goed.	
8	Die roomys was pienk		20	Die vrou het 'n trui aangehad.	
10	Hy het sy vinger gesny.		151	Hy het laat by die huis gekom.	
110	Die polisieman soek 'n hond.		220	Die mes is vol botter.	
		SNR50:			SNR50:
	Lys 3			Lys 4	
21	Die kamer word nou koud.		31	Hulle hardloop verby die huis.	
22	Die vrou het haar man gehelp.		32	Die trein het ontspoor.	
23	Die ou man is bekommerd.		33	Hy het by die venster uitgeval.	
24	'n Seuntjie hardloop in die pad af.		34	Die park is naby die pad.	
26	Hy het sy boetie gekry.		35	Die kok het uie gesny.	
27	Hulle staan op hulle knieë.		36	Die hond gee 'n kwaai grom.	
28	Die meisie het haar pop verloor.		37	Iemand het die geld gevat.	
29	Die kind gryp die speelding.		38	My pa kom huis toe.	
30	Die vuurhoutjies lê op die rak.		39	Die lorrie ry in die straat af.	
173	Hy loer oor die muurtjie.		40	Die slim dogtertjies lees boek.	
		SNR50:			SNR50:
	Lys 5			Lys 6	
41	Die vrou het haar huis opgeruim.		51	Die gesin het 'n huis gekoop.	
42	Die hond het teruggekem.		52	Die beker staan op die rak.	
43	Die vrugte lê op die grond.		53	Hulle het gaan kaas koop.	
44	Die bus het vroeg gery.		54	Sy skryf vir haar boetie 'n brief.	
45	Hulle het twee leë bottels.		55	Die speler het die bal verloor.	
46	Die bal het gehop.		56	Die meisies luister musiek.	
47	My pa het die brood vergeet.		57	Die boek vertel 'n storie.	
62	Hulle is weg met vakansie.		58	Sy het naby haar venster gestaan.	
112	Hulle het na die prent gestaar.		60	Die vyf mans werk hard.	
117	Hulle het aan die venster geklop.		94	Hy het sy geld laat val.	
		SNR50:			SNR50:
	Lys 7			Lys 8	
50	Die nuwe pad is op die kaart.		71	Die grond was te hard.	
61	Hy luister na sy pa.		72	Die emmers is vol water.	
63	Die trein beweeg vinnig.		73	Die hoender het eiers gelê.	
64	Die kar het in 'n muur vasgejaag.		74	Die bestuurder wag op die hoek.	
65	Die skoonmaker gebruik 'n besem.		75	Die polisieman ken die pad.	
66	Sy het in die spieël gekyk.		76	Die seuntjie klim in die bed.	
67	Hulle het die paadjie gevolg.		77	Die twee boere gesels lekker.	
68	Die hond spring op die stoel.		78	Ma het blomme gepluk.	
69	Hy het die brief gaan pos.		79	Die voortuin lyk baie mooi.	
70	Die melk staan op die tafel.		80	Hy het sy hoed verloor.	
		SNR50:			SNR50:

	Lys 9			Lys 10	
81	Die krane is bokant die wasbak.		59	Die tafel het drie pote.	
82	Pa het by die hek betaal.		91	Die seun het swart hare.	
83	Ons het gaan brood koop.		92	Die vragmotor ry teen die bult op.	
84	Die wedstryd is verby.		93	Die ou vrou was by die huis.	
85	Sy dra 'n klomp inkopiesakkies.		95	Hulle het al die eiers gebreek.	
86	Die seun het 'n rooi karretjie.		96	Sy help haar maatjie.	
87	Hulle het 'n uur lank gewag.		98	Die gras word nou lank.	
88	Die groot hond is gevaarlik.		99	Die vuur was baie warm.	
89	Die aarbeikonfyt was soet.		100	Hy suig nog sy duim.	
90	Die plant staan langs die deur.		165	Hy kruip agter die bos weg.	
	SNR50:			SNR50:	
	Lys 11			Lys 12	
9	Die leer staan by die deur.		101	Die seuntjie hardloop skool toe.	
48	Die meisie het 'n inkleurboek.		111	Die bestuurder het verdwaal.	
102	Daar is oulike mense wat kom.		113	Die oond se deur was oop.	
103	Daar groei blomme in die tuin.		114	Die kar ry baie vinnig.	
104	Die klein babatjie is mooi.		115	Die verwer het 'n kwas gebruik.	
105	Die dogters het tafel gedek.		116	Hy drink uit sy beker.	
106	Hulle het oor die gras geloop.		118	Die skêr is nogal skerp.	
107	Hy het sy oë toegemaak.		119	Sy bel haar dogter.	
108	Hulle het die ambulans gebel.		120	Die hond het die kat gejaag.	
109	Sy betaal vir die brood.		149	Ons hond is baie siek.	
	SNR50:			SNR50:	
	Lys 13			Lys 14	
121	Hulle hou van appelkooskonfyt.		131	Die skoonmaker vee die vloer.	
122	Sy ma het die venster toegemaak.		132	Die badwater was warm.	
123	Hy speel buite saam met sy maatjie.		133	Hy probeer die lepel bykom.	
124	Die wolke bring reën.		134	Hy het sy rekening betaal.	
125	Hulle het die muur geverf.		135	Die vadoek is nogal nat.	
126	Die handdoek het op die vloer geval.		136	Die meisie het verkoue gekry.	
127	Die hond eet 'n stuk vleis.		137	Die twee kinders lag.	
128	Die reën val op die dak.		138	Die peperpot was leeg.	
129	Die gesin eet graag vis.		139	Die hond het uit 'n bak gedrink.	
130	Suiker is baie soet.		140	'n Meisie kom by die deur in.	
	SNR50:			SNR50:	
	Lys 15			Lys 16	
49	Die lemoen was nogal soet.		17	Die klein babatjie slaap.	
141	Die pad loop teen die bult op.		153	Sy pak die mandjie vol kos.	
142	Piesangs is geel vrugte.		154	Daar is 'n mier op sy voet.	
143	Die koei lê op die gras.		155	Sy skryf haar naam op die bord.	
144	Hy het sy sussie bang gemaak.		156	Die wolke gaan reën bring.	
145	My pa het druiwe gepluk.		157	Hy klim op tot bo.	
146	Die ketel het vinnig gekook.		158	Daar was baie min mense.	
147	Hulle het geld verloor.		159	Die paleis het 'n pragtige tuin.	
148	Sy skep dit met 'n lepel.		160	Die visstok se katrol is stukkend.	
150	Hy het 'n fiets geleen.		211	Sy lees 'n dik boek.	
	SNR50:			SNR50:	

	Lys 17			Lys 18	
97	Die bordjie wys die pad aan.		25	Die huis het 'n mooi tuin.	
161	Hy blaas die stof van sy kas af.		171	Ons was gister biblioteek toe.	
163	Die kinders groet die juffrou.		172	Die blaartjie dryf in die stroom af.	
164	Sy het haar elmboog gestamp.		174	Sy plak 'n seël op die brief.	
166	Hulle gaan na die wedstryd kyk.		175	Die weerlig slaan hard.	
167	Sy buk om haar tas op te tel.		176	Die konstabel groet vriendelik.	
168	Hy trek 'n sirkel om die woord.		177	Die seuns is baie lui.	
169	Die bal het hom teen die kop getref.		178	Die hond vee sy snoet aan my af.	
170	Daar is 'n geraamte in die kis.		180	Die vrou dra baie juwele.	
205	Hy maak die boot met 'n tou vas.		217	Ons moet oor die brug stap.	
	SNR50:			SNR50:	
	Lys 19			Lys 20	
181	Die polisieman is gewapen.		191	Die seuntjie vee die stoep.	
182	Die meisie het sproete op haar neus.		192	Die hondjie se pels blink mooi.	
183	Die kar se ratte maak 'n geraas.		193	Hy is uitgeput na die wedstryd.	
184	Hulle het die vakansie gaan ski.		194	Sy pluk 'n rooi roos.	
185	Die vrou is deftig aangetrek.		195	Hulle steek 'n kers op.	
186	Die dogtertjie wil 'n ponie hê.		196	Sy spring oor die muurtjie.	
187	Die brood is van graan gemaak.		197	My pa plant 'n boom.	
188	Daar is 'n swerm bye by die nes.		198	n By het my sussie gestee.	
189	Die bank is gister beroof.		199	Ons moet vroeg in die bed klim.	
190	Sy streel haar pop se hare.		200	Ons soek die pad op die kaart.	
	SNR50:			SNR50:	
	Lys 21			Lys 22	
162	Sy was gister by die haarkapper.		18	Die hond het met 'n stok gespeel.	
201	Die stoel staan in die hoek.		152	Iemand het die boek by my geleen.	
202	Die hond jaag die kat.		179	Ek en my pa speel skaak.	
203	Hy koop 'n lamp vir sy bedkassie.		212	Die vlag wapper in die wind.	
204	Die hasie sit in sy hok.		213	Hy het sy been gebreek.	
206	Die vrou kom by die winkel uit.		214	My boonste knoop het afgeval.	
207	Sy kam haar pop se hare.		215	Sy vurk het op die vloer geval.	
208	Hy het kaas en melk gaan koop.		216	Hulle speel buite met die bal.	
209	Ons eet pap en wors vanaand.		218	Hy praat met sy mond vol kos.	
210	Die bome se blare val af.		219	Sy pa het 'n bok gaan skiet.	
	SNR50:			SNR50:	

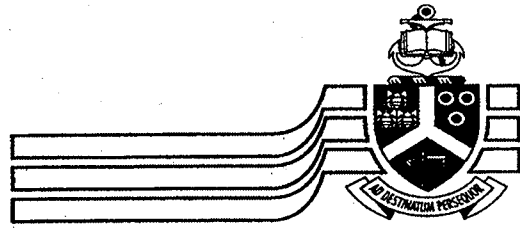
APPENDIX H:
TEST FORM – PHONETICALLY BALANCED LISTS

	Lys 1			Lys 2	
4	My pa sluit die voorhek.		10	Hy het sy vinger gesny.	
30	Die vuurhoutjies lê op die rak.		53	Hulle het gaan kaas koop.	
36	Die hond gee 'n kwaai grom.		72	Die emmers is vol water.	
54	Sy skryf vir haar boetie 'n brief.		77	Die twee boere gesels lekker.	
73	Die hoender het eiers gelê.		82	Pa het by die hek betaal.	
90	Die plant staan langs die deur.		86	Die seun het 'n rooi karretjie.	
158	Daar was baie min mense.		92	Die vragmotor ry teen die bult op.	
159	Die paleis het 'n pragtige tuin.		98	Die gras word nou lank.	
171	Ons was gister biblioteek toe.		110	Die polisieman soek 'n hond.	
208	Hy het kaas en melk gaan koop.		197	My pa plant 'n boom.	
	SNR50			SNR50	
	Lys 3			Lys 4	
41	Die vrou het haar huis opgeruim.		23	Die ou man is bekommerd.	
43	Die vrugte lê op die grond.		26	Hy het sy boetie gekry.	
61	Hy luister na sy pa.		33	Hy het by die venster uitgeval.	
102	Daar is oulike mense wat kom.		66	Sy het in die spieël gekyk.	
105	Die dogters het tafel gedek.		101	Die seuntjie hardloop skool toe.	
152	Iemand het die boek by my geleen.		103	Daar groei blomme in die tuin.	
174	Sy plak 'n seël op die brief.		125	Hulle het die muur geverf.	
175	Die weerlig slaan hard.		156	Die wolke gaan reën bring.	
189	Die bank is gister beroof.		167	Sy buk om haar tas op te tel.	
221	Sy het die stoof aan vergeet		177	Die seuns is baie lui.	
	SNR50			SNR50	
	Lys 5			Lys 6	
39	Die lorry ry in die straat af.		8	Die roomys was pienk	
40	Die slim dogtertjies lees boek.		20	Die vrou het 'n trui aangehad.	
58	Sy het naby haar venster gestaan.		42	Die hond het teruggekom.	
64	Die kar het in 'n muur vasgejaag.		70	Die melk staan op die tafel.	
107	Hy het sy oë toegemaak.		89	Die aarbeikonfyt was soet.	
121	Hulle hou van appelkooskonfyt.		133	Hy probeer die lepel bykom.	
138	Die peperpot was leeg.		137	Die twee kinders lag.	
139	Die hond het uit 'n bak gedrink.		145	My pa het druiwe gepluk.	
153	Sy pak die mandjie vol kos.		149	Ons hond is baie siek.	
217	Ons moet oor die brug stap.		186	Die dogtertjie wil 'n ponie hê.	
	SNR50			SNR50	
	Lys 7			Lys 8	
12	Kersfees is in die somer.		2	Sy sny met 'n mes.	
14	Die muis hardloop na sy gat toe.		17	Die klein babatjie slaap.	
37	Iemand het die geld gevat.		49	Die lemoen was nogal soet.	
50	Die nuwe pad is op die kaart.		51	Die gesin het 'n huis gekoop.	
91	Die seun het swart hare.		59	Die tafel het drie pote.	
116	Hy drink uit sy beker.		136	Die meisie het verkoue gekry.	
151	Hy het laat by die huis gekom.		140	'n Meisie kom by die deur in.	
176	Die konstabel groet vriendelik.		142	Piesangs is geel vrugte.	
181	Die polisieman is gewapen.		195	Hulle steek 'n kers op.	
215	Sy virk het op die vloer geval.		207	Sy kam haar pop se hare.	
	SNR50			SNR50	

	Lys 9			Lys 10	
1	Die hanswors het 'n snaakse gesig.		18	Die hond het met 'n stok gespeel.	
13	Die polisie het die kar gejaag.		29	Die kind gryp die speelding.	
16	Daar is 'n hoop hout onder die boom.		34	Die park is naby die pad.	
55	Die speler het die bal verloor.		62	Hulle is weg met vakansie.	
71	Die grond was te hard.		80	Hy het sy hoed verloor.	
124	Die wolke bring reën.		93	Die ou vrou was by die huis.	
126	Die handdoek het op die vloer geval.		109	Sy betaal vir die brood.	
184	Hulle het die vakansie gaan ski.		114	Die kar ry baie vinnig.	
188	Daar is 'n swerm bye by die nes.		219	Sy pa het 'n bok gaan skiet.	
203	Hy koop 'n lamp vir sy bedkassie.		222	Die mot vlieg al om die lig.	
		SNR50			SNR50
	Lys 11			Lys 12	
11	Die seun het die speletjie geken.		3	Die huis het nege kamers.	
24	'n Seuntjie hardloop in die pad af.		9	Die leer staan by die deur.	
27	Hulle staan op hulle knieë.		60	Die vyf mans werk hard.	
65	Die skoonmaker gebruik 'n besem.		67	Hulle het die paadjie gevolg.	
69	Hy het die brief gaan pos.		81	Die krane is bokant die wasbak.	
94	Hy het sy geld laat val.		119	Sy bel haar dogter.	
111	Die bestuurder het verdwaal.		127	Die hond eet 'n stuk vleis.	
183	Die kar se ratte maak 'n geraas.		130	Suiker is baie soet.	
211	Sy lees 'n dik boek.		169	Die bal het hom teen die kop getref.	
218	Hy praat met sy mond vol kos.		196	Sy spring oor die muurtjie.	
		SNR50			SNR50
	Lys 13			Lys 14	
19	Hulle sê 'n klomp lawwe goed.		6	Die sak sleep op die grond.	
32	Die trein het ontspoor.		46	Die bal het gehop.	
76	Die seuntjie klim in die bed.		78	Ma het blomme gepluk.	
84	Die wedstryd is verby.		115	Die verwer het 'n kwas gebruik.	
88	Die groot hond is gevaarlik.		131	Die skoonmaker vee die vloer.	
128	Die reën val op die dak.		178	Die hond vee sy snoet aan my af.	
162	Sy was gister by die haarkapper.		185	Die vrou is deftig aangetrek.	
164	Sy het haar elmboog gestamp.		190	Sy streek haar pop se hare.	
179	Ek en my pa speel skaak.		192	Die hondjie se pels blink mooi.	
180	Die vrou dra baie juwele.		193	Hy is uitgeput na die wedstryd.	
		SNR50			SNR50
	Lys 15			Lys 16	
7	Die seun loop op sy hande.		25	Die huis het 'n mooi tuin.	
95	Hulle het al die eiers gebreek.		96	Sy help haar maatjie.	
112	Hulle het na die prent gestaar.		99	Die vuur was baie warm.	
157	Hy klim op tot bo.		129	Die gesin eet graag vis.	
165	Hy kruip agter die bos weg.		141	Die pad loop teen die bult op.	
168	Hy trek 'n sirkel om die woord.		147	Hulle het geld verloor.	
187	Die brood is van graan gemaak.		148	Sy skep dit met 'n lepel.	
191	Die seuntjie vee die stoep.		161	Hy blaas die stof van sy kas af.	
199	Ons moet vroeg in die bed klim.		201	Die stoel staan in die hoek.	
204	Die hasie sit in sy hok.		206	Die vrou kom by die winkel uit.	
		SNR50			SNR50

	Lys 17			Lys 18	
22	Die vrou het haar man gehelp.		38	My pa kom huis toe.	
35	Die kok het uie gesny.		44	Die bus het vroeg gery.	
63	Die trein beweeg vinnig.		45	Hulle het twee leë bottels.	
97	Die bordjie wys die pad aan.		74	Die bestuurder wag op die hoek.	
113	Die oond se deur was oop.		85	Sy dra 'n klomp inkopiesakkies.	
194	Sy pluk 'n rooi roos.		117	Hulle het aan die venster geklop.	
198	n By het my sussie gesteek.		118	Die skêr is nogal skerp.	
202	Die hond jaag die kat.		122	Sy ma het die venster toegemaak.	
205	Hy maak die boot met 'n tou vas.		172	Die blaartjie dryf in die stroom af.	
216	Hulle speel buite met die bal.		214	My boonste knoop het afgeval.	
	SNR50			SNR50	
	Lys 19			Lys 20	
5	Hulle kyk na die horlosie.		21	Die kamer word nou koud.	
15	Die vrou maak 'n speelding.		48	Die meisie het 'n inkleurboek.	
52	Die beker staan op die rak.		100	Hy suig nog sy duim.	
57	Die boek vertel 'n storie.		106	Hulle het oor die gras geloop.	
120	Die hond het die kat gejaag.		132	Die badwater was warm.	
123	Hy speel buite saam met sy maatjie.		134	Hy het sy rekening betaal.	
135	Die vadoek is nogal nat.		150	Hy het 'n fiets geleen.	
144	Hy het sy sussie bang gemaak.		163	Die kinders groet die juffrou.	
209	Ons eet pap en wors vanaand.		200	Ons soek die pad op die kaart.	
220	Die mes is vol botter.		210	Die bome se blare val af.	
	SNR50			SNR50	
	Lys 21			Lys 22	
31	Hulle hardloop verby die huis.		28	Die meisie het haar pop verloor.	
47	My pa het die brood vergeet.		56	Die meisies luister musiek.	
75	Die polisieman ken die pad.		68	Die hond spring op die stoel.	
79	Die voortuin lyk baie mooi.		83	Ons het gaan brood koop.	
87	Hulle het 'n uur lank gewag.		108	Hulle het die ambulans gebel.	
104	Die klein babatjie is mooi.		146	Die ketel het vinnig gekook.	
155	Sy skryf haar naam op die bord.		154	Daar is 'n mier op sy voet.	
160	Die visstok se katrol is stukkend.		166	Hulle gaan na die wedstryd kyk.	
170	Daar is 'n geraamte in die kis.		212	Die vlag wapper in die wind.	
182	Die meisie het sproete op haar neus.		213	Hy het sy been gebreek.	
	SNR50			SNR50	

APPENDIX B: INFORMED CONSENT LETTER



Universiteit van Pretoria

Departement Kommunikasiepatologie
Spraak- Stem- en Gehoorkliniek

Tel : +27 12 420 2303

Faks : +27 12 420 3517

Navorser: Marianne Theunissen

Tel No: 082 696 8869

Faks no: 012 998 4076

E-pos adres: theunissenm@gmail.com

Geagte Deelnemer

Ek is 'n nagraadse student aan die Universiteit van Pretoria, en is tans besig met 'n studie wat die ontwikkeling van 'n **Afrikaanse toets vir die bepaling van spraakherkennings-drempels in geraas deur die gebruik van sinsmateriaal** behels. Aan die hand van hierdie studie sal 'n navorsingsverhandeling saamgestel word wat ingedien sal word as deel van die vereistes vir 'n Meestersgraad in Kommunikasiepatologie.

Waaroor handel die studie?

Die onderskeiding van spraak, veral in die teenwoordigheid van agtergrondsgeraas, is 'n uiters belangrike deel van alledaagse kommunikasie. Die evaluasie van hierdie vaardigheid is dus van groot belang vir oudioloë wêreldwyd. Die gebruik van sinsmateriaal in so 'n evaluasie blyk ook 'n baie nuttige instrument te wees, aangesien sinne verteenwoordigend van gesprekspraak is en dus 'n goeie aanduiding kan gee van 'n persoon se alledaagse gehoorvermoëns. In Suid-Afrika, meer spesifiek in Afrikaans, is daar tans geen spesifieke metode om hierdie vaardigheid te evalueer nie. In die lig van hierdie situasie, asook internasionale ontwikkelings van soortgelyke materiaal, blyk dit dat die ontwikkeling van so 'n toets in Afrikaans tans 'n uitvoerbare en waardevolle proses sal wees. **Die hoofdoel van**

hierdie studie is dus die ontwikkeling en evaluasie van 'n Afrikaanse toets van sinsherkenningdrempels in geraas.

Hoe gaan die data ingesamel word?

Die ontwikkeling van hierdie toets gaan verskeie stappe behels. Nadat die sinsmateriaal saamgestel is, gaan die ***natuurlikheid van die sinne deur Afrikaanssprekendes geëvalueer word*** (Groep A van die deelnemers). Die sinne gaan gevolglik digitaal opgeneem word en 'n gepaste agtergrondsgeraas gaan gegenereer word. Hierna gaan verskeie vrywilligers (Afrikaanssprekend, met normale gehoor – Groep B) ***gevra word om na die sinne te luister*** in die teenwoordigheid van bepaalde hoeveelhede agtergrondsgeraas. Die luisteraars moet die sinne (of dele van die sinne) wat hulle hoor hardop herhaal. Deur 'n proses van herhaalde sifting gaan 'n bepaalde hoeveelheid sinne met 'n soortgelyke moeilikheidsgraad geselekteer word en in lyste gegroepeer word om die finale toets saam te stel.

Wat gaan van my verwag word?

U word vriendelik versoek om die aangehegte ***ingeligte toestemmingsvorm te onderteken***. Indien u in Groep A van die deelnemers is, gaan daar van u verwag word om 'n lys van 520 kort, Afrikaanse ***sinne te lees en te gradeer in terme van die natuurlikheid daarvan***. Indien u in groep B van die deelnemers is, sal u 'n ***siftingsgehoortoets*** ondergaan waartydens bepaal sal word of u gehoordrempels binne normale perke val. Daarna sal u gevra word om in 'n klankdigte kamer deur oorfone te ***luister na 'n aantal sinne (soms in die teenwoordigheid van agtergrondsgeraas)***. Daar gaan vereis word dat u die ***sinne hardop herhaal*** aan die persoon wat die toets administreer. Die totale duur van die prosedure gaan afhang van die fase waarin die navorsing is, maar kan wissel van 45 minute tot 2 ure. U behoort nie enige ongemak tydens die prosedure te verduur nie, aangesien die intensiteit van die klanke wat u gaan hoor (die sinne sowel as die geraas) teen gemiddelde of sagte intensiteite aangebied sal word.

Wat gaan gebeur met die data wat versamel is?

Al die inligting wat tydens die studie verkry word sal as ***streng vertroulik en konfidensieel*** hanteer word. Geen individu se naam sal genoem word in die verhandeling of die wetenskaplike artikel wat aan die einde van die studie saamgestel gaan word nie. Die data gaan ook vir 'n tydperk van vyftien jaar in elektroniese formaat gestoor word, maar sal steeds as konfidensieel hanteer word.

Wat moet ek nou doen?

Teken asseblief die aangehegte toestemmingsvorm. Daarna sal die navorser u inlig omtrent verdere reëlins en prosedures.

Wat kry ek hieruit?

U deelname aan die studie sal **bydra tot die ontwikkeling van 'n unieke audiologiese toets in Suid-Afrika**, wat die potensiaal het om in die toekoms gebruik te word om **dienslewering aan persone met gehoorprobleme te verbeter**. Indien u in Groep B van die deelnemers is, sal u 'n **gratis gehoorsiftingsstoets** ondergaan.

Wat as ek van plan verander?

Indien u op 'n stadium besluit dat u nie meer aan die studie wil deelneem nie, mag u op enige tydstop onttrek sonder enige gevolge aangesien **deelname volkome vrywilliglik** is.

Waar en hoe kan ek die resultate sien?

Die **resultate van u persoonlike gehoorsiftingsstoets kan op u versoek aan u verskaf word**. 'n Verwerkte weergawe van al die data wat ingesamel word, sal gepubliseer word in 'n gepaste wetenskaplike joernaal na afloop van die studie. U kan die navorser kontak (per telefoon of e-pos) vir 'n kopie van hierdie artikel.

Indien u enige vrae het of enige van die aspekte soos hierbo verduidelik onduidelik vind, rig asseblief u vrae aan die navorser. Byvoorbaat dankie vir u waardevolle deelname.

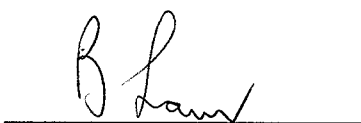
Die uwe,



Marianne Theunissen
Navorser/Oudioloog



Dr De Wet Swanepoel
Studieleier



Prof Brenda Louw
Hoof: Departement Kommunikasiepatologie

-
- Statistics South Africa. (2004). *Provincial Profile 2004. Gauteng*. Retrieved January 23, 2008 from <http://www.statssa.gov.za/>
- Statistics South Africa. (2007). *Mid-year population estimates 2007*. Retrieved July 7, 2007 from <http://www.statssa.gov.za/>
- Stockdell, K.G. (1980). Measuring Discrimination Efficiency: Classical Approaches. In R.R. Rupp & K.G. Stockdell, Sr (Eds.), *Speech Protocols in Audiology* (pp. 99-116). New York: Grune & Stratton Inc.
- Stockley, K.B. & Green, W.B. (2000). Interlist equivalency of the Northwestern University Auditory Test No. 6 in quiet and noise with adult hearing-impaired individuals. *Journal of the American Academy of Audiology*, 11(2), 91-96.
- Strom, K.E. (2003). The HR 2003 Dispenser Survey. *Hearing Review Online*, 10(6). Retrieved January 22nd, 2008, from http://www.hearingreview.com/issues/articles/2003-06_02.asp.
- Struwig, F.W. & Stead, G.B. (2001). *Planning, designing and reporting research*. Cape Town: Pearson Education South Africa.
- Stuart, A., Phillips, D.P. & Green, W.B. (1995). Word recognition performance in continuous and interrupted broad-band noise by normal-hearing and simulated hearing-impaired listeners. *The American Journal of Otology*, 16(5), 658-663.
- The Design-Based Research Collective. (2003). Design-Based Research: An Emerging Paradigm for Educational Inquiry. *Educational Researcher*, 32(1), 5-8.
- Tobias, J. (1964). On phonemic analysis of speech discrimination tests. *Journal of Speech and Hearing Research*, 7, 99-100.

- Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., et al. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *International Journal of Audiology*, 44(6), 358-369.
- Van Wieringen, A. & Wouters, J. (2006). LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and The Netherlands. In preparation for the *International Journal of Audiology*.
- Ventry, I.M. & Schiavetti, N. (1980). *Evaluating Research in Speech Pathology and Audiology: A Guide for Clinicians and Students*. Ontario: Addison-Wesley Publishing Company.
- Versfeld, N.J., Daalder, L, Festen, J.M. & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *Journal of the Acoustical Society of America*, 107(3), 1671-1684.
- Wagener, K.C. (2004). Factors Influencing Sentence Intelligibility in Noise. DSc Thesis. Oldenburg: BIS-Verlag. Retrieved January 24, 2008 from <http://docserver.bis.uni-oldenburg.de/publikationen/dissertation/2003/wagfac03/pdf/wagfac03.pdf>
- Wagener, K.C. & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *International Journal of Audiology*, 44(3), 144-156.
- Wang, F. & Hannafin, M.J. (2005). Design-Based Research and Technology-Enhanced Learning Environments. *Educational Technology Research and Development*, 53(4), 5-23.

- White, S.C. (1980). Assessment of Sensory and/or Peripheral-Neural Lesion Sites. In R.R. Rupp & K.G. Stockdell, Sr (Eds.), *Speech Protocols in Audiology* (pp. 145-162). New York: Grune & Stratton Inc.
- Wiley, T.L. & Fowler, C.G. (1997). *Acoustic Immittance Measures in Clinical Audiology: A Primer*. San Diego: Singular Publishing Group, Inc.
- Wilson, R.H. & Margolis, R.H. (1983). Measurements of auditory thresholds for speech stimuli. In D.F. Konkle & W.F. Rintelmann (Eds.), *Principles of Speech Audiometry* (pp 79-126). Baltimore: University Park Press.
- Wilson, R.H., Zizz, C.A., Shanks, J.E. & Causey, G.D. (1990). Normative Data in Quiet, Broadband Noise, and Competing Message for Northwestern University Auditory Test no. 6 by a Female Speaker. *Journal of Speech and Hearing Disorders*, 55, 771-778.
- Wilson, R.H. & Strouse, A. (1999). Psychometrically Equivalent Spondaic Words Spoken by a Female Speaker. *Journal of Speech, Language and Hearing Research*, 42, 1336-1346.
- Wilson, R.H. & Carter, A.S. (2001). Relation Between Slopes of Word Recognition Psychometric Functions and Homogeneity of the Stimulus Materials. *Journal of the American Academy of Audiology*, 12(1), 7-14.
- Wilson, R. H. & McArdle, R. (2005). Speech signals used to evaluate functional status of the auditory system. *Journal of Rehabilitation Research and Development*, 42(4), 79-94.
- Wilson, R.H., Carnell, C.S. & Cleghorn, A.L. (2007). The Words-In-Noise (WIN) Test with Multitalker Babble and Speech-Spectrum Noise Maskers. *Journal of the American Academy of Audiology*, 18(6), 522-529.
- Wilson, R.H., McArdle, R.A. & Smith, S.L. (2007). An Evaluation of the BKB-SIN, HINT, QuickSIN, and WIN Materials on Listeners With Normal Hearing

and Listeners With Hearing Loss. *Journal of Speech, Language, and Hearing Research*, 50, 844-856.

Wong, L.L.N & Soli, S.D. (2005). Development of the Cantonese Hearing In Noise Test. *Ear & Hearing*, 26(3), 276-289.

Wong, L.L.N., Soli, S.D., Liu, S., Han, N. and Huang, M. (2007). Development of the Mandarin Hearing in Noise Test (MHINT). *Ear and Hearing*, 28(2), Supplement, 70S-74S.