Modernising speech audiometry: using a smartphone application to test word recognition

Marianne van Zyl[1], De Wet Swanepoel*[1,2,3,4], Hermanus Carel Myburgh[5]

1. Department of Speech-Language Pathology and Audiology, University of Pretoria, South Africa
2. Callier Center for Communication Disorders, University of Texas, Dallas, USA,
3. Ear Sciences Centre, School of Surgery, University of Western Australia, Nedlands, Australia
4. Ear Science Institute Australia, Subiaco, Australia
5. Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa

***Corresponding author: De Wet Swanepoel**

Postal address: Department of Speech-Language Pathology and Audiology, Room 3-5, University of Pretoria, Cnr Lynnwood and University Roads, Pretoria, South Africa, 0002

Telephone number: +27 12 420 4280

Email: dewet.swanepoel@up.ac.za

Acronyms/Abbreviations:

ANOVA	Analysis of variance
App	Application
CD	Compact Disc
CID	Central Institute for the Deaf
CV	Consonant-Vowel
CVC	Consonant-Vowel-Consonant
dB HL	Decibel Hearing Level
FFT	Fast Fourier Transform
FVEWA	Foneties Verteenwoordigende Eenlettergrepige Woordlyste in Afrikaans
ISI	Interstimulus intervals
mHealth	Mobile health
MLV	Monitored Live Voice

NH	Normal-hearing
PI function	Performance-intensity function
PTA	Pure tone average
RMSE	Root-mean-square-error
SRT	Speech reception threshold
VU-meter	Volume units meter

Abstract

Objective: This study aimed to develop and assess a method to measure word recognition abilities using a smartphone application (App) connected to an audiometer. *Design:* Word lists were recorded in South African English and Afrikaans. Analyses were conducted to determine the effect of hardware used for presentation (computer, compact-disc player, or smartphone) on the frequency content of recordings. An Android App was developed to enable presentation of recorded materials via a smartphone connected to the auxiliary input of the audiometer. Experiments were performed to test feasibility and validity of the developed App and recordings. *Study sample:* Participants were 100 young adults (18-30 years) with pure tone thresholds ≤ 15 dB across the frequency spectrum (250-8000 Hz). *Results:* Hardware used for presentation had no significant effect on the frequency content of recordings. Listening experiments indicated good inter-list reliability for recordings in both languages, with no significant differences between scores on different lists at each of the tested intensities. Performance-intensity functions had slopes of 4.05%/dB for English and 4.75%/dB for Afrikaans lists at the 50% point. *Conclusions:* The developed smartphone App constitutes a feasible and valid method for measuring word recognition scores, and can support standardisation and accessibility of recorded speech audiometry.

Keywords

Speech perception, tele-audiology, mobile health, word recognition, speech audiometry

Introduction

Speech audiometry is an essential part of the basic audiological test battery. However, the diagnostic value of speech audiometry depends on the use of reliable and standardised procedures and materials (Roeser & Clark, 2008). The use of recorded speech materials has long been recognised as a more reliable method than monitored live voice (MLV), with a mounting body of supporting evidence (see e.g. Hood & Poole, 1980; Mendel & Owen 2011; Mullenix et al, 1989; Penrod, 1979; Uhler et al, 2016). Brandy (1966) reported that presentations of the same list by the same speaker on different days can result in significant differences of nearly 10% in listener performance, while Penrod (1979) reported differences of up to 38% in scores between different talkers presenting the same test materials.

Despite this evidence, the last audiometric practice survey conducted among audiologists in the United States indicated that a large majority (82%) of clinicians in the United States of America continued to use MLV for speech audiometry (Martin et al, 1998). Unfortunately, no recent surveys are available to indicate whether this situation has improved (Mendell & Owen, 2011). Reasons cited for the use of MLV include savings in time and costs, increased flexibility and patient performance, as well as the ability to use a local accent for testing (Roeser & Clark, 2008). In a survey among South African audiologists with regard to speech audiometry practices, all of the respondents (n = 84) used MLV for word recognition testing (Roets, 2006). The majority of respondents (62.7%) felt that recorded materials would give them less control over the test situation, and 66.2% agreed with the statement that the presenters of existing recordings usually had a foreign accent. In many contexts like South Africa, there appears to be a need for locally recorded tests that are accessible and widely

available, and that offers clinicians more control over the test situation than existing recordings that are available on audiocassette or compact disc (CD).

Among the clinicians surveyed in the Martin et al (1998) survey, only 1% used digitised speech, while 12% used CDs and 4% used tape recordings. According to Roeser and Clark's (2008) informal survey, costs and set-up of equipment are some of the reasons why clinicians prefer the use of MLV over recorded speech. In addition, the fixed inter-stimulus intervals (ISIs) on CD recordings can cause frustration for the clinician and the listener. Recently developed computer-based audiometers offer the option of embedded recorded speech materials that provide accuracy and flexibility (Mikolai & Mroz, 2010). However, this option may not be affordable to all audiologists as it requires procurement of a new audiometric equipment.

There is, however, a possible alternative in technology that could improve accessibility and affordability of recorded speech materials for audiometry. The Economist (2015) estimates that there are 2 billion people using smartphones worldwide and that by 2020, around 80% of adults will own a smartphone. The wide distribution, mobility and fluidity of smartphones make them an ideal platform for a variety of uses, including telehealth, or more specifically mobile health (mHealth). A number of recent publications have reported on the successful use of smartphone applications in audiology (Mahomed-Asmail et al, 2015; Potgieter et al, 2016; Sandström et al, 2016; Swanepoel et al, 2014; Yousuf-Hussein et al, 2016). Smartphones have the capability to present digital recordings, to track and display test information, and can be connected to an audiometer via the auxiliary input, making them a possible platform to facilitate speech audiometry tests. In light of the promising capabilities of smartphones as audiometric tools, and the need for widely accessible, reliable and flexible

testing options for speech audiometry, the present study endeavoured to test the feasibility of a smartphone application (App) as a platform for word recognition testing.

The main objective of this study was to develop and assess the feasibility and validity of an App to measure word recognition abilities by connecting a smartphone to an audiometer. As a secondary objective, word lists suitable to the research context were developed and evaluated in terms of validity and reliability.

The project was conducted in three phases, namely i) word list preparation and evaluation; ii) word list recordings and frequency analyses; and iii) evaluation of the lists in normal-hearing (NH) listeners using an App. The method and results of each phase are described and discussed below.

Phase I: Preparation and evaluation of word lists

Monosyllabic word lists suitable for word recognition testing were prepared in (South African) English and Afrikaans. Although only 9.6% of South Africans use English as a home language (Statistics South Africa, 2011), it is considered the lingua franca (Hanekom, 2014) and many South Africans have a working knowledge of English. Afrikaans is reported to be the home language of 13.5% of South Africans, the third most common home language after isiZulu and isiXhosa (Statistics South Africa, 2011). The majority of South African audiologists are native speakers of English or Afrikaans (Roets, 2006) and receive their training in English (Swanepoel, 2006). Because word recognition testing requires the test administrator to judge the correctness of a listener's response, the present study focused on the implementation of Afrikaans and English as test languages, to enable widespread use of the developed materials and App by local audiologists.

Materials and methods

The Afrikaans word lists that were used are called "Foneties Verteenwoordigende Eenlettergrepige Woordlyste in Afrikaans" (FVEWA), that is, phonetically representative monosyllabic word lists in Afrikaans (unpublished). Based on the phoneme frequencies reported in Van Heerden (1999) the same audiologist subsequently developed the lists. Based on a relatively recent (1999) sample of spoken Afrikaans, the lists are phonemically representative of Afrikaans, and are phonemically matched or balanced across lists. The FVEWA consists of six lists that are each 25 words long. Because these lists were recently developed in South Africa, they were not adapted or revised for the present study.

The CID W-22 word lists, developed in the 1950's (Hirsh et al, 1952) are the most commonly used standardised English lists in South Africa (Roets, 2006). According to Hirsh et al (1952), the words in these lists were rated to be familiar by a panel of five judges, with the words "ace", "ale", and "pew" considered to be of doubtful familiarity. The familiarity of the words used in a speech recognition test, can affect the content validity of the test (Ostergard, 1983). If some of the words in the lists are unfamiliar to a listener, the results of the test are more likely to be a reflection of their vocabulary than their ability to accurately perceive meaningful monosyllabic words. In light of the fact that the CID-W22 lists were developed in the United States, and the vocabulary in these lists has not been reviewed for its familiarity since the 1950's, these lists were adapted for the present study in a two-step process.

The first step included reviewing lists to identify words that might be unfamiliar to South African listeners in the present day, based on the researchers' knowledge and experience of South African English. Possible alternatives for the potentially unfamiliar words were selected in a manner that would preserve the phonetic balancing of the lists. For example, in

List 2, the word “ail” was considered to be a potential problem. The word “aim” was selected as a possible alternative. To replace the removed (alveolar lateral approximant) phoneme /l/ and in order to keep the same number of (bilabial nasal) /m/ phonemes in the list, the word “dumb” was replaced with the word “dull”. It should be noted that the original CID W-22 lists were balanced in terms of syllable types (e.g. consonant-vowel (CV), consonant-vowel-consonant (CVC) etc.). However, pronunciation of the lists with a South African English accent, which is much closer to British than American pronunciation, altered syllable structure in many instances. The production of an “r” at the end of a word as pronounced with an American accent, for example, is replaced with an elongated vowel in British and South African accents, changing syllable structure from CVC to CV in words like “there” and “where”. For this reason, syllable structure was not taken into account in the adaptation of the lists, which rather focused on familiarity and phonetic balance between lists.

The second step in the preparation of the word lists was to submit the vocabulary to a familiarity survey conducted among two groups of participants. The first group consisted of native South African English speakers (n = 36; 25 females). This group was selected using purposive sampling to compile a group representing a wide range of age groups (19 – 78 years) and professions in an attempt to provide a representative sample of the native English speakers in the South African population. Participants were provided with the words in the CID W-22 lists, along with the additional words that were selected as possible replacements for potentially problematic words, and were instructed to read through the words carefully and indicate if there were any words that were not familiar to them (defined on the survey as words that they do not know the meaning of). The second group of participants consisted of audiologists with clinical experience with the CID-W22 word lists (n = 10). The limited sample size was due to the fact that the majority of audiologists in South Africa use untitled

word lists obtained from academic institutions (Roets, 2006), and therefore do not have experience with standardised word lists such as the CID-W22. The participating audiologists were asked to indicate words in the CID W-22 lists that were problematic for their patients, by using one of three codes (1, 2 or 3). Coding a word with "1" indicated that patients often have difficulty recognising this word, even at suprathreshold levels where all or most other words in the list are recognised correctly. The code "2" indicated that patients reported or the audiologist suspected that the word was unfamiliar to them. . In cases where the audiologist felt that both the descriptions of "1" and "2" applied (i.e. a word that was often misheard and was suspected to be unfamiliar to some patients), the word was coded with "3".

Words that were considered unfamiliar by any number of native speakers, as well as words that were considered problematic by a majority of the participating audiologists (>5/10) were excluded from the adapted lists.. This method was followed because the survey among audiologists had some limitations. Firstly, their rating was a subjective opinion of their patients' experience of the lists, and was based on a population that included second- or third-language speakers of English. In addition, because MLV is used exclusively by the responding audiologists as presentation method, all the audiologists' experiences with the lists were based on their own pronunciation of the words. Incidentally, seven out of the 10 audiologists in the survey speak English as a second language. Therefore, if a minority of audiologists rated the word as problematic, it might have been due to their own pronunciation of the word.

Results and discussion

A total of eight words were removed from the four CID-W22 lists based on the familiarity surveys. There was good correspondence between the responses by audiologists and native speakers (Table 1). All of the words considered unfamiliar by native speakers were also rated

as problematic by the majority (>5) of audiologists. Only three words (bathe, owes and ought) that were not marked as unfamiliar by native speakers were removed from the original lists because they were rated as problematic by the majority of audiologists. All the words that were added to the lists to replace problematic or unfamiliar words were included in the survey among native speakers, with the exception of “bake” and “love”. Both of these words appear in the Longman Communication 3000 word list, a list of the 3000 most frequent English words, based on statistical analysis of the 390-million-word Longman Corpus Network (Bullon & Leech, 2007).

Table 1: Words removed from the lists, showing the number of native speakers that identified the word as unfamiliar, along with the number of audiologists who rated the word as problematic

Word	List no.	No. of native speakers who report this word to be unfamiliar (n = 36)	No. of audiologists who identifies this word as problematic (n = 10)
mew	1	14	9
bathe	1	0	6
ail	2	10	6
pew	2	11	9
tare	2	16	7
owes	3	0	9
darn	4	5	8
ought	4	0	8

Appendix A (published at http://tandfonline.com/doi/suppl.) shows the new English lists, along with words that were removed and added in order to improve the familiarity of the lists while retaining phonetic balance. The final collection consisted of four lists of 50 words each.

Phase II: Word list recordings and frequency analyses

Following the digital recording of the word lists, frequency analyses were conducted on the recorded materials as a means to assess the validity of the smartphone App as a presentation method. The content validity (Ostergard, 1983) of a speech audiometry test could be affected not only by the words in the test, but also by its frequency content. If, for example, recorded speech was presented using a device with a limited frequency bandwidth, the frequency

content of the presented speech would not correspond to the original recordings, and therefore would not be measuring a response to the content it was intended to measure. To determine the content validity of the speech tests as conducted through a smartphone, Fast Fourier Transform (FFT) analyses were conducted on the original recordings in digital waveform format and compared to FFT analyses of the recordings played back via CD-player, laptop and two different smartphones. At present, laptops and CD-players are the only hardware options available for playing recorded speech materials through an audiometer, in cases where digital audiometers with embedded recordings are not used. Differences in frequency distributions between a specific method of presentation and the original waveform could indicate reduced validity of the method, as the presented materials would not be an accurate representation of the original recording, and would in reality be measuring perception of a "filtered" version of the speech the test was intending to present.

Method

Word lists were digitally recorded in a professional recording studio with double-walled soundproofing, using a Røde NT1-A 1" cardioid condenser microphone with a frequency range of 20 Hz – 20 kHz. The microphone was positioned on a microphone stand 20 cm from the speaker's mouth. The Afrikaans lists were produced by a female audiologist whose voice had previously been evaluated for articulation, resonance and voice quality. This was the same speaker who produced the sentences for the Afrikaans test of sentence recognition thresholds in noise (Theunissen et al, 2011). The English lists were produced by a female audiologist who is a native speaker of South African English and speaks with an accent that is considered to be representative of the province of South Africa where the research was conducted (Gauteng). Recorded .wav files were edited using *Praat* (Boersma & Weenink, 2016), to remove unwanted silences, leaving 100 ms of silence before and after each utterance. Any unwanted sounds or artefacts in the recordings were removed, and the

intensity (root-mean-square or rms value) of each utterance was subsequently modified to ensure equal intensity of all utterances.

To determine the influence of the hardware used for presentation of the materials, frequency analyses were performed. For all analyses the recorded materials were normalised with respect to the original (digital) material files' energy content, while the original (digital) material files' energy contents were normalised to unity. A subset of fifteen utterances from the Afrikaans word lists were purposefully selected to represent all the phonemes of the test, and to include utterances from each of the six recorded lists. These utterances were recorded onto a laptop (Dell XPS L502X, Windows 7 Home Premium 64-bit, Intel i7-2670QM, 6 GB RAM) using a standard auxiliary cable connecting the source and the laptop with 9mm male jacks on both ends. Recordings were made from a Samsung J2 smartphone, a Samsung Trend Neo smartphone, a laptop (Asus K5410U, Windows 10 64-bit, Intel Core i5-7200U, 8GB RAM) and a CD-player (Sansui CD-210) using a CD containing the original materials. After time- and frequency energy normalisation, the root mean squared error (RMSE) was calculated between the recorded and the original digital material for their frequency representations.

Results and discussion

Results of the frequency comparisons between the original recordings in digital format and the recordings played through a CD-player, laptop, and two smartphones are summarised in Table 2.

Table 2: Root mean squared errors (RMSEs) between FFTs of original recordings and recordings presented through two smartphones (J2 and Trend Neo), laptop, and CD-player

	FFT RMSE: original versus…			
	J2	Trend	Laptop	CD-player
blits	0.0027827	0.0013114	0.0062525	0.0198000
bring	0.0020354	0.0042163	0.0063188	0.0040170
deel	0.0036644	0.0017855	0.0070343	0.0014666
droog	0.0019334	0.0007440	0.0059403	0.0008657
erg	0.0041547	0.0008121	0.0100020	0.0033379
hemp	0.0027627	0.0028171	0.0092617	0.0024900
hier	0.0018641	0.0032497	0.0071479	0.0009909
hond	0.0021630	0.0041677	0.0069458	0.0017379
jag	0.0023993	0.0019357	0.0007551	0.0024943
maak	0.0033533	0.0008039	0.0067243	0.0003763
moeg	0.0021156	0.0015096	0.0004389	0.0002863
stout	0.0016073	0.0048067	0.0031780	0.0136030
stry	0.0029071	0.0019943	0.0040467	0.0069882
vrug	0.0008888	0.0007451	0.0018011	0.0148070
was	0.0076912	0.0017824	0.0072573	0.0003811
Average	**0.00282**	**0.00218**	**0.00554**	**0.00491**

Due to the sample size, a non-parametric Friedman's ANOVA was conducted on these findings to determine the significance of the difference between the RMSE's of the different presentation methods. The ANOVA indicated that the difference between the four methods was not significant, $\chi^2(3) = 7.80$, $p = 0.05$.

Results of the frequency analyses indicated that the hardware used for playback of the digital recordings did not have a significant effect on the frequency content of the recordings. This finding supports the use of low-cost smartphones as a platform for the presentation of recorded word lists, as the frequency content of the original recording is sufficiently retained and comparable to that delivered through a CD-player or laptop.

Phase III: Evaluation of recorded word lists using a smartphone App in normal-hearing listeners

The third phase of the project evaluated the feasibility of the App as a platform for measuring word recognition scores. This was achieved by connecting a smartphone to the audiometer and determining whether it was possible to attain sufficient intensity for accurate calibration, and to present, score, and plot word recognition results. The duration of the test was also measured and compared to existing methods reported in the literature as a measure of feasibility.

During this phase, the reliability, validity, and sensitivity of the word lists developed and recorded in the present work were also evaluated. This was achieved by presentation of the developed lists using the App connected to an audiometer in a sample of young, NH adults. Results of the tests were used to evaluate inter-list reliability as a measure of the coefficient of equivalence (Ostergard, 1983), psychometric slopes as an indication of test sensitivity (Theunissen et al, 2009), and inter-listener variability as an indirect measure of specificity. Criterion-related validity (Ostergard, 1983) was assessed by comparing maximum recognition scores to scores reported in literature for NH listeners.

Research participants

One hundred NH listeners with ages ranging from 18 to 30 years participated in the study. All listeners underwent pure tone audiometry to determine pure tone averages (PTAs) and to ensure that pure tone thresholds were < 20 dB HL at octave frequencies from 250 to 8000 Hz, Forty participants who listened to the English lists reported that English was their native language (language most often spoken at home), while 60 participants who listened to the Afrikaans lists were native Afrikaans speaking.

Software and test setup

A custom-made Android App called "hearSpeech" was developed to serve as a platform for the presentation of the developed recordings. To conduct testing using this App, a Samsung Galaxy J2 smartphone was plugged into the auxiliary input of the audiometer with a 3.5 mm male jack in the smartphone to a stereo RCA connector on a clinical audiometer (GSI 61). The audiometer served as the attenuator through which the presentation intensity was controlled. The App featured a calibration function which played a 1000 Hz calibration tone referenced to the same intensity as the test words. While the calibration tone was playing, the volume of the external channel where the smartphone was plugged in was adjusted until the VU-meter of the audiometer reached zero. Initially, the intensity of the recordings played through the smartphone was too low. The recorded words were subsequently re-scaled to a higher intensity (75 dB) using *Praat*, which resulted in successful calibration (VU-meter reaching zero).

Details about each listener (name, age, gender) were saved in the App along with test results and duration. The App offered the user (audiologist) a choice between Afrikaans and English as test language. The user could also choose which list to present to a listener, and could manually enter the intensity and test ear as selected on the audiometer.

Procedure

For the duration of the test, listeners were seated in a double-walled sound booth, with the audiologist outside the booth in a sound-treated room. Once details were entered about the listener, test setup (intensity, test ear, test language and test list) was selected, and calibration was completed, the App provided instructions to the user for administering the test, as well as instructions to be read out to the listener. After the start button was pressed, the words in the selected list were presented one by one. During presentation, a text version of the test word

was presented on the screen of the smartphone, along with two buttons – one indicating that the response was correct and the other to indicate an incorrect response (Figure 1). Once the listener repeated the test word, the test administrator made a judgment about the correctness of the response and touched the corresponding button, and the next word was automatically presented. The App offered the option of using a carrier phrase as well as adjusting the delay before the word was played to anything between 0 and 1000 ms. In the present work, Afrikaans listeners were tested with a 200 ms delay and a carrier phrase. After the first 15 participants were tested in Afrikaans, it was observed that the App's ability to present the next word after a listener's response made the use of a delay before presentation unnecessary. For this reason, and because the English lists were longer, English listeners were tested without a delay and carrier phrase in order to limit the test time. If selected, the App could plot a performance-intensity function at the end of the test.

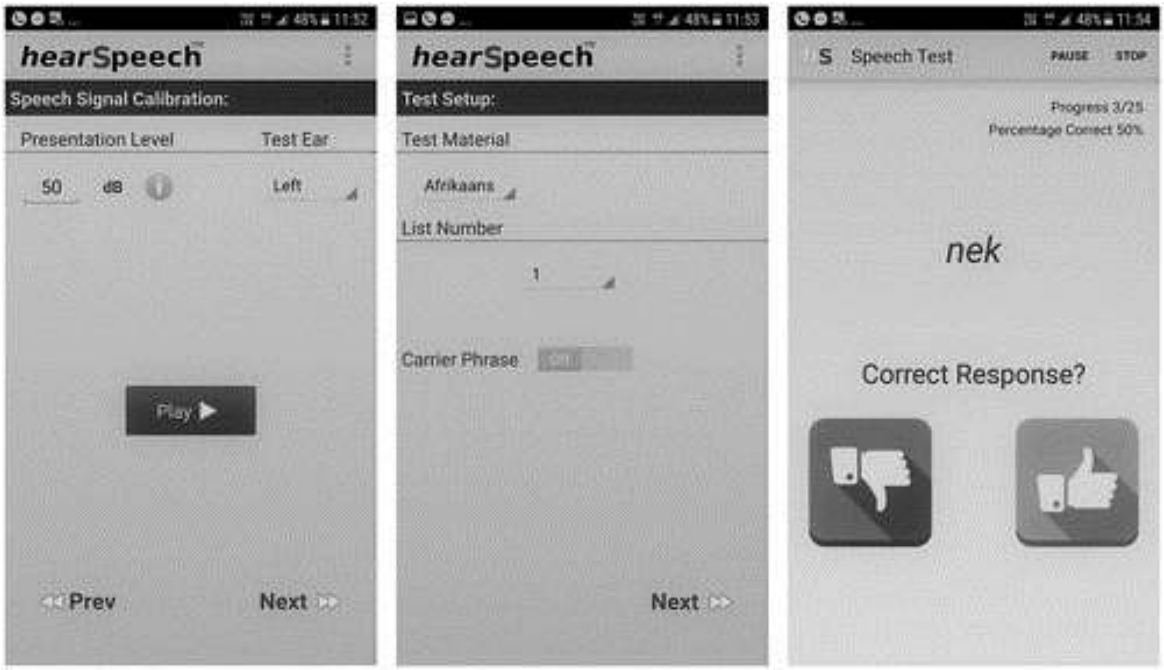

Figure 1: Illustration of hearSpeech App screens, from left to right: calibration, test setup, response tracking.

To validate the English lists, 40 NH listeners were tested. Each listener was presented all four lists, each at a different intensity. The Afrikaans lists were validated in 60 NH listeners, each

listening to all six lists, with each list presented at a different intensity. Test order and list intensity were counterbalanced between subjects. Table 3 shows the test order and intensities for Afrikaans and English tests. Each group consisted of 10 listeners. Each listener was tested monaurally, using their best ear. The pure tone average (PTA, average of thresholds at 500, 1000 and 2000 Hz) was calculated for each listener's best ear, and presentation levels were determined accordingly. For example, if a listener's PTA was 5 dB, and they were in Group 1 for the Afrikaans test, List 1 was presented at 10 dB (5dB + PTA), List 2 at 15 dB (10 dB + PTA) etc. Although the speech reception threshold (SRT) is often used as a reference to determine intensity levels for word recognition testing, no recorded materials are available in Afrikaans or South African English to reliably measure SRTs. Since PTAs and SRTs are expected to be in good agreement in NH listeners (Brandy, 2002; Preece & Fowler, 1992), the PTA was used as a reference in the present work.

Table 3: Test order and intensities of lists presented. Intensities shown are in dB re:PTA

AFRIKAANS	**30dB**	**25dB**	**20dB**	**15dB**	**10dB**	**5dB**
Group 1	List6	List5	List4	List3	List2	List1
Group 2	List5	List4	List3	List2	List1	List6
Group 3	List4	List3	List2	List1	List6	List5
Group 4	List3	List2	List1	List6	List5	List4
Group 5	List2	List1	List6	List5	List4	List3
Group 6	List1	List6	List5	List4	List3	List2
ENGLISH	**35dB**	**25dB**	**15dB**	**5dB**		
Group 1	List4	List3	List2	List1		
Group 2	List1	List4	List3	List2		
Group 3	List2	List1	List4	List3		
Group 4	List3	List2	List1	List4		

Results and discussion

The present work demonstrated the feasibility of using a smartphone App to assess word recognition ability. It was found that connection of a smartphone to an audiometer was possible and adequate intensity levels for calibration of the test signal could be reached. It was also demonstrated that the word recognition score could be measured and plotted using

the smartphone App developed for this purpose.

The word recognition results obtained from NH listeners on the Afrikaans word lists (n = 60) and English word lists (n = 40) are summarised in Table 4.

Standard deviations from the mean scores at different intensities were reported for each of the evaluated lists (Table 4). This gives an indication of the variability between different listeners' scores on each of the lists. It may also provide an indirect measure of specificity, as large variability among a NH population could mean that NH listeners may achieve scores that deviate a great deal from the mean score of the NH group, which could cause over-referral. As could be expected, variability was greater at lower presentation intensities, where listeners would have relied on guessing when presented words could not be heard clearly. At the highest intensity tested (30 dB above PTA for Afrikaans lists and 35 dB above PTA for English lists), standard deviations ranged from 1.75 to 2.71% for English lists and between 2.07 and 4.24% for the Afrikaans lists. The larger variability as shown by the larger standard deviations of the Afrikaans lists can be explained by the shorter length of these lists (25 words as compared to the 50-word English lists). As demonstrated by the binomial model of Thornton and Raffin (1978), shorter lists result in larger variability in recognition scores. The standard deviations for the 50-word English lists of the current study were similar to standard deviations reported for the NU-6 word lists, which ranged from 1.0 to 3.8% at 32 dB above SRT (Tillman & Carhart, 1966). In light of the finding in the present study that a 50-word list took, on average, only 30 seconds longer to administer than a 25-word list, it may be advisable to use two of the 25-word Afrikaans lists per intensity to increase test reliability. Combining two full-length lists would maintain phonetic representation and phonemic balance.

Table 4: Average performance (%) of NH listeners on Afrikaans word lists at six intensities (n = 10 per intensity) and English lists at four intensities (n = 10 per intensity)

English

Presentation intensity		**List 1**			**List 2**			**List 3**			**List 4**	
in dB re:PTA	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
5	25	27.4	17.9	26	24.8	8.9	24	28.2	19.2	25	27.0	14.8
15	66	64.6	17.9	76	71.0	21.4	58	59.0	12.6	76	74.6	14.5
25	97	95.6	6.0	90	90.0	6.7	93	92.6	3.8	92	90.8	5.6
35	98	98.2	2.4	98	97.0	2.7	99	98.4	2.1	98	98.2	1.8

Afrikaans

Presentation intensity		**List 1**			**List 2**			**List 3**			**List 4**			**List 5**			**List 6**	
in dB re:PTA	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
5	22.5	23.3	16.9	10	10.8	6.5	16	14.8	11.3	16	18.4	17.7	22	23.6	10.6	30	28.0	17.7
10	58	52.8	24.1	42	44.0	22.6	32	27.6	10.2	28	30.0	12.7	40	46.4	19.0	48	48.6	16.3
15	74	78.0	13.2	66	64.4	17.9	72	67.2	21.5	64	64.4	9.7	64	62.4	19.2	68	67.2	17.9
20	90	88.0	10.3	90	88.4	8.1	90	86.8	14.0	84	82.4	11.7	82	77.6	11.7	84	82.4	14.5
25	96	94.0	6.9	96	93.2	6.5	96	96.0	3.3	96	96.0	4.2	96	90.0	17.0	96	94.4	3.4
30	98	97.2	3.8	100	97.2	4.2	98	97.2	3.8	100	98.0	3.4	100	98.4	2.8	100	98.4	2.1

The steepness of the performance-intensity slope is thought to be an indication of the homogeneity of the individual test items (Wilson & Carter, 2001), as well as the sensitivity of the test (Theunissen et al, 2009). The slope of the graph between the 20% and 80% points was calculated in accordance with the traditional linear model, which assumes a linear relation between these two points (Wilson & Carter, 2001). For the Afrikaans lists, this slope was 4.38 %/dB (see Figure 2). At the 50% point on the graph, the slope was 4.75 %/dB. The intensity at which the average score across listeners and lists was 50% was 11.64 dB above their PTA for the Afrikaans lists. The performance-intensity slope of the Afrikaans lists is similar to slope values reported for monosyllabic word lists in other languages, such as 2.8-4.2 %/dB for Spanish (Flores & Aoyama, 2008), 4.1 and 3.47%/dB for Mandarin (Han et al, 2009; Wu et al, 2011), 3.45-3.53%/dB for Arabic (Garadat et al, 2017), 6.2 %/dB for Turkish (Durankaya et al, 2014) and 4.52-4.64%/dB for Telugu, a South Indian Dravidian language (Kumar & Mohanty, 2012).

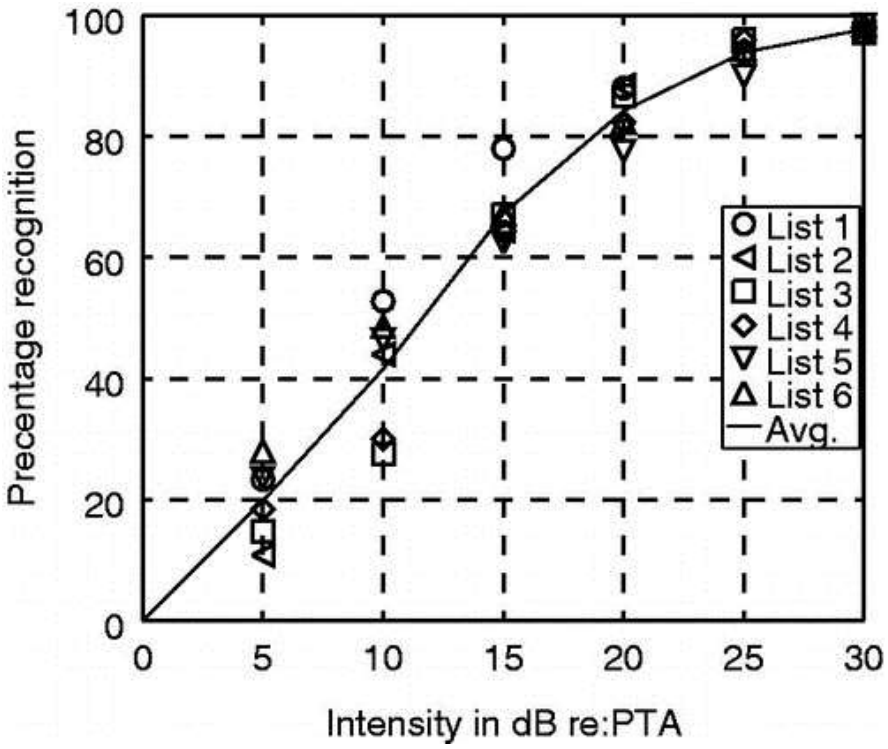

Figure 2: Recognition scores for Afrikaans word lists across listeners (n = 10 per list at each intensity) as a function of presentation intensity. The line indicates the average PI function of the six lists.

Performance on the individual lists were compared across listeners at each of the intensity levels tested using Friedman's ANOVA. Results indicated that there was no significant difference between performances on lists at all of the tested intensities, except at 10 dB, χ^2 (5) = 13.71, $p < 0.05$. However, post hoc Wilcoxon pairwise comparisons with Bonferroni corrections applied did not find significant differences between any of the lists.

The average slope for the English lists between the 20% and 80% points on the graph is 3.59 %/dB, and 4.05 %/dB at the 50% point (see Figure 3). The intensity at which the average score across listeners and lists was 50% was 10.72 dB above their PTA. Slope values reported in the literature for different recordings of the original CID-W22 lists are 4.0 %/dB (Flores & Aoyama, 2008), 4.9%/dB (Beattie et al, 1985) and 4.1%/dB (Heckendorf et al, 1997). The slope value of 3.59%/dB found in the current study is therefore comparable to previous reports.

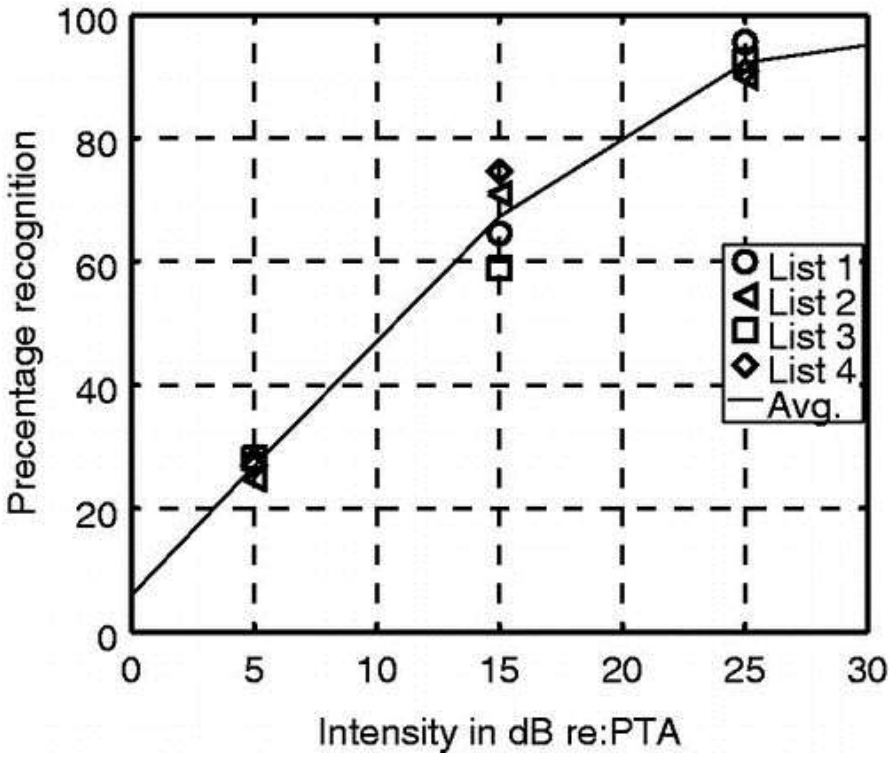

Figure 3: Recognition scores for English word lists across listeners (n = 10 per list at each intensity) as a function of presentation intensity. The line indicates the average PI function of the four lists.

Performance on the individual lists were compared across listeners at each of the intensity levels tested using Friedman's ANOVA. Results indicated that there was no significant difference between performances on lists at all of the tested intensities, except at 25 dB, $\chi^2(3) = 9.32$, $p < 0.05$. However, post hoc Wilcoxon pairwise comparisons with Bonferroni corrections applied did not find significant differences between any of the lists.

The results obtained from young NH listeners in the present study indicated that both the Afrikaans and English lists showed a high degree of inter-list equivalence, as demonstrated by the absence of significant differences in recognition scores between individual lists at all of the tested intensities. Inter-list equivalence is an important form of reliability in speech audiometry (Ostergard, 1983), as small differences between scores on different lists would indicate different lists can be used on the same listener without significantly influencing the test results (Theunissen et al, 2009).

To evaluate the criterion-related validity (Ostergard, 1983) of the developed lists, the intensities at which the NH listeners in the present study obtained a maximum score were compared to typical values reported in the literature. Maximum scores on a speech recognition task are typically achieved at 30-40 dB above the speech reception threshold (SRT) in NH listeners (Brandy, 2002; McArdle & Hnath-Chisolm, 2015). In the present work, SRTs were not measured, as no recorded test materials were available. However, the SRT is expected to closely correspond to the three-frequency PTA that was used as a reference in the present work (Brandy, 2002). The levels at which maximum scores were obtained in the present work with the English and Afrikaans lists therefore correspond well to levels reported for NH listeners in existing literature.

As an additional measure of feasibility, the duration of the smartphone test method was

evaluated. According to Mendell and Owen (2011), one of the reasons why clinicians reportedly prefer the less reliable method of MLV over recorded speech materials is that MLV is considered to be quicker. In the present study, average test duration across listeners and lists, including the 200 ms delay and the carrier phrase was 109 seconds (1 minute 49 seconds) per list for the 25-word Afrikaans lists. If the duration of the carrier phrase was deducted, average duration would be 106 seconds, and if the 200 ms delay per word was deducted, average duration would be 101 seconds. This is longer than MLV (just under one minute), similar to short ISI CD presentations (1min30s), but shorter than the long ISI recordings (approximately 2min15s) of 25-word (half) lists reported by Mendell and Owen (2011).

For the English lists in the present work, each 50 words long and presented without a carrier phrase or stimulus delay, the average test duration per list across listeners and lists was 132 seconds (2 minutes and 12 seconds). This was slightly longer than MLV presentations of 50-word lists reported by Mendell and Owen (2011), which was just under two minutes. It was shorter than presentation times with a CD, which resulted in test times of between three and five minutes, depending on the ISI (Mendell & Owen, 2011). Differences in test duration of less than a minute per list are not considered clinically significant (Mendell & Owen, 2011). Therefore, the use of the developed App to present recorded word lists resulted in test administration times similar to MLV and CD presentations, with the added advantage of offering a flexible method of timing the presentation according to the listener's responses.

Contributions and limitations

The results of the present work indicate that a smartphone can be used for reliable and valid assessment of word recognition using the developed smartphone App. This method offers a

test platform for audiologist that is more reliable than the prevalent MLV method, but has more flexibility than the use of CD-recordings with fixed ISIs. It also does not require the purchase of additional expensive equipment such as a digital audiometer or a CD-player that can be connected to the audiometer. The use of a smartphone App to present recorded word lists for speech recognition testing offers the additional advantage of ease of distribution amongst audiologists. The App could be downloaded from an App store, which can be done remotely, without any delays or additional costs that may be incurred by delivery of a physical product such as a CD or digital audiometer. Cost structure for the App has not yet been finalised, but will ensure that the App provides a cost-effective platform for speech audiometry. Future improvements of the App (such as additional test languages, speech tests or user options/features) can be implemented with minimal effort by downloading updates from the App store. Having established the feasibility of using the smartphone as a test platform for speech audiometry, future work could include development of tests to measure SRTs, uncomfortable loudness levels, speech recognition in noise, and many others using the App-based platform. An iOS version of the App will also be developed in the future.

To reduce test time, English listeners were tested without the use of a delay and a carrier phrase. There are conflicting reports in existing literature on whether a carrier phrase affects word recognition test results (see Brandy, 2002 for a discussion). Many audiologists use a carrier phrase to help them present words at the correct intensity when using MLV (Brandy, 2002), and this is unnecessary when using recorded materials. However, recent work has recommended the use of a carrier phrase when measuring aided word recognition, due to slow attack times in amplitude compression of some hearing aids (Versfeld & Goverts, 2013). Future work should explore the effect of a carrier phrase in word recognition testing

when using recorded materials in a flexible test setup such as offered by the App, or a digital audiometer.

In the present work, the use of a smartphone App as a platform for word recognition testing was evaluated using newly developed monosyllabic word lists with no previously published data on their reliability. The use of previously validated recordings may have simplified the evaluation of the App in the present work, as it would have reduced the number or new variables introduced and enabled comparison to other test platforms (e.g. CD or MLV) using the same lists. However, the lack of such data on South African word lists necessitated the development and recording of new lists. If the results indicated a lack of validity or reliability it may have been difficult to disentangle the effects of the smartphone platform and that of the newly recorded lists. Since the results demonstrated good reliability and validity, it appears that both the developed lists and the smartphone platform were shown to be suitable for clinical use. The adaptation and recording of the monosyllabic word lists in South African English and Afrikaans offer the additional advantage of thoroughly validated lists that may help to improve current speech audiometry methods in South Africa.

Test-retest reliability was not evaluated in the present work. The superiority of recorded speech materials over MLV in terms of test-retest reliability has long been established (Mendell & Owen, 2011), and since the use of the smartphone App is merely a different platform to present recorded materials, there was no reason to expect differences in test-retest reliability as compared to other methods of presenting recorded speech.

Conclusion

A smartphone App was developed that enables the presentation of recorded monosyllabic word materials to test speech recognition, using the audiometer as an attenuator to control intensity levels. Monosyllabic word lists were recorded in South African English and Afrikaans to enable implementation and evaluation of the App. The use of the smartphone App was shown to be feasible and valid, and the developed lists were shown to be valid and reliable measures of monosyllabic word recognition.

Declaration of interest statement

The second and third authors have a relationship with the hearX Group, who owns the right to the intellectual property, which includes equity, consulting and potential royalties.

References

Beattie R. & Raffin M. 1985. Reliability of threshold, slope, and PB Max for monosyllabic words. *J Speech Hear Disord, 50,* 166-178 .

Boersma P. & Weenink D. 2016. Praat: doing phonetics by computer [Computer program]. Version 6.0.21, retrieved 25 September 2016 from http://www.praat.org/

Brandy W.T. 1966. Reliability of voice tests of speech discrimination. *J Speech Hear Res,* 9, 461 – 465.

Brandy, W. 2002. Speech audiometry. In J. Katz (ed.) Handbook of Clinical Audiology: Fifth edition. Philadelphia: Lippincott Williams & Wilkins, pp. 96–110.

Bullon S., & Leech G. 2007. Longman Communication 3000 and the Longman Defining Vocabulary. In Longman Communication 3000. (pp. 1-7). Harlow: Pearson Longman. Available at http://www.lextutor.ca/freq/lists_download/longman_3000_list.pdf. Last accessed on 4 November 2016.

Durankaya S.M., Şerbetçioğlu B., Dalkılıç G., Gürkan S. & Kırkım, G. 2014. Development of a Turkish monosyllabic word recognition test for adults. *Int Adv Otol 10*(2), 172-180.

Flores L. & Aoyama K. 2008. A comparison of psychometric performance on four modified Spanish word recognition tests. *Texas J Audiol Speech-Lang Path, 31*, 64-70.

Garadat S.N., Abdulbaqi K.J. & Haj-Tas M.A. 2017. The development of the University of Jordan word recognition test. Int J Audiol. 56(6), 424-430.

Han D., Wang S., Zhang H., Chen J., Jiang W., Mannell R., Newall P. & Zhang L. 2009. Development of Mandarin monosyllabic speech test materials in China. *Int J Audiol. 2009, 48*(5), 300-311.

Hanekom T.H. 2014. Comparison of the South African spondaic wordlist and the CID W-1 for measuring speech recognition threshold. Unpublished M Communication Pathology (Audiology) dissertation. University of Pretoria.

Heckendorf A.L., Wiley T.L. & Wilson R.H. 1997. Performance norms for the VA compact disc versions of CID W-22 (Hirsh) and PB-50 (Rush Hughes) word lists. *J Am Acad Audiol, 8*(3), 163-172.

Hirsh I.J., Davis H., Silverman S.R., Reynolds E.G., Eldert E. et al. 1952. Development of materials for speech audiometry. *J Speech Hear Dis,* 17(3), 321 – 337.

Hood J.D. & Poole J.P. 1980. Influence of the speaker and other factors affecting speech intelligibility. *Audiology , 19,* 434 – 455.

Kumar S.B.R. & Mohanty P. 2012. Speech recognition performance of adults: a proposal for a battery for Telugu. *Theory Practice Lang Studies, 2(2)*, 193-204.

Mahomed-Asmail F., Swanepoel D., Eikelboom R.H., Myburgh H.C. & Hall J. 3rd. 2015. Clinical validity of hearScreen™ smartphone hearing screening for school children. *Ear Hear*, 37(1), e11-e17.

Martin F.N., Champlin C.A. & Chambers J.A. 1998. Seventh survey of audiometric practices in the United States. *J Am Acad Audiol* , 9, 95 – 104.

McArdle R. & Hnath-Chisolm T. 2015. Speech audiometry. In *Handbook of Clinical Audiology*, J. Katz, ed., Wolters Kluwer, 61-75.

Mendel L.L. & Owen S.R. 2011. A study of recorded versus live voice word recognition. *Int J Audiol, 50*(10), 688-693.

Mikolai, T. & Mroz, A.C. 2010. Modern speech audiometry with integrated recorded speech materials. *The Hearing Review*. Available at www.hearingreview.com/2010/11/modern-speech-audiometry-with-integrated-recorded-speech-materials/. Last accessed June 2nd, 2017.

Mullennix J.W., Pisoni D.B. & Martin C.S. 1989. Some effects of talker variability on spoken word recognition. *J Acoust Soc Am , 85*(1), 365– 378.

Ostergard C.A. 1983. Factors influencing the validity and reliability of speech audiometry. *Seminars Hear 4*(3), 221-239.

Penrod J.P. 1979. Talker effects on word-discrimination scores of adults with sensorineural hearing impairment. *J Speech Hear Dis , XLIV*, 340 – 349.

Potgieter J., Swanepoel D., Myburgh H.C., Hopper T.C. & Smits, C. 2016. Development and validation of a smartphone-based digits-in-noise hearing test in South African English. *Int J Audiol*, 55(7), 405-411.

Preece J.P. & Fowler C.G. 1992. Relationship of pure-tone averages to speech reception threshold for male and female speakers. *J Am Acad Audiol 3*, 221-224.

Roeser R.J. & Clark J.L. 2008. Live voice speech recognition audiometry: Stop the madness! *Audiology Today , 20*(1), 32 – 33.

Roets, R. 2006. Spraakoudiometrie in Suid-Afrika: ideale kriteria teenoor kliniese praktyk. Unpublished MCommunication Pathology dissertation, University of Pretoria.

Sandström J., Swanepoel D., Myburgh H.C. & Laurent C. 2016. Smartphone threshold audiometry in underserved primary health-care contexts. *Int J Audiol*, 55(4), 232-238.

Statistics South Africa. 2011. Census 2011 key results, Statistics South Africa, Pretoria.

Swanepoel, D. 2006. Audiology in South Africa. *Int J Audiol*, 45(5), 262-266.

Swanepoel D., Myburgh H.C., Howe D.M., Mahomed F. & Eikelboom R.H. 2014. Smartphone hearing screening with integrated quality control and data management. *Int J Audiol*, 53(12), 841-849.

The Economist. 2015. Telecoms and society: The truly personal computer. Available at: http://www.economist.com/news/briefing/21645131-smartphone-defining-technology-age-truly-personal-computer 28 February 2015-last update. [last accessed 2016, October].

Theunissen M., Swanepoel, D. & Hanekom, J. 2009. Sentence recognition in noise: Variables in compilation and interpretation of tests. *Int J Audiol,48*, 743-757.

Theunissen M., Hanekom J.J. & Swanepoel D. (2011). The development of an Afrikaans test for sentence recognition thresholds in noise. *Int J Audiol, 50,* 77-85.

Thornton A.R. & Raffin M.J. 1978. Speech-discrimination scores modeled as a binomial variable. *J Speech Hear Res, 21*(3), 507-518.

Tillman, T.W. & Carhart, R. 1966. An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University auditory test no. 6. *Technical report number SAM-TR-66-55*. US Air Force School of Aerospace Medicine, Aerospace Medical Division, Brooks AFB.

Uhler K., Biever A. & Gifford R.H. 2016. Method of Speech Stimulus Presentation Impacts Pediatric Speech Recognition: Monitored Live Voice Versus Recorded Speech. *Otology & Neurotology 37*:e70–e74

Van Heerden R. (1999). *Die voorkomsfrekwensie van die spraakklanke van Afrikaans met die oog op fonetiese balansering van oudiologie woordelyste*, B Communication Pathology dissertation, thesis, Department of Communication Pathology, University of Pretoria.

Versfeld N.J. & Goverts S.T. 2013. The effect of a carrier phrase on hearing aid amplification of single words in quiet. *Int J Audiol, 52,* 189-193.

Wilson R.H. & Carter A.S. 2001. Relation between slopes of word recognition psychometric functions and homogeneity of the stimulus materials. *J Am Acad Audiol, 12*, 7-14.

Wilson W.J. & Moodley S. 2000. Use of the CID W22 as a South African English speech discrimination test. *S Afr J Commun Disord, 47,* 57-62.

Wu W., Zhang H., Chen J., Chen J. & Lin C. 2011. Development and evaluation of a computerized Mandarin speech test system in China. *Comput Biol Med 41,* 131–138.

Yousuf-Hussein S., Swanepoel D., Biagio de Jager, L., Myburgh, H.C., Eikelboom, R.H. & Hugo, J. 2016. Smartphone hearing screening in mHealth assisted community-based primary care. *J Telemed Telecare*, 22(7), 405-412.

Appendix A – CID-W22 lists adapted according to survey results. New words are indicated in italics, removed words are listed at the bottom of each list.

List1	**List2**	**List3**	**List4**
ace	*aim*	add	aid
an	air	aim	all
as	and	are	am
bake	*are*	ate	arm
bells	*bear*	bill	at
carve	by	book	*barn*
chew	cap	camp	be
could	cars	chair	*can*
dad	chest	cute	chin
day	die	do	clothes
deaf	does	done	cook
earn	*dull*	dull	*dart*
east	ease	ears	dolls
few	eat	end	dust
give	else	*eye*	ear
high	flat	farm	eyes
him	gave	glove	few
hunt	ham	hand	go
isle	hit	have	hang
it	hurt	he	his
jam	ice	if	in
knees	ill	is	jump
law	jaw	jar	leave
low	key	king	men
me	knee	knit	my
melt	live	*love*	near
none	move	may	net
not	new	nest	*nut*
or	now	no	of
owl	odd	oil	our
poor	off	on	pale
ran	one	out	*red*
see	own	pie	save
she	*poke*	raw	shoe
skin	rooms	say	so
stove	send	*shows*	*sought*
them	show	smooth	stiff
there	smart	start	tea
they	*stew*	tan	than
thing	that	ten	they
toe	then	this	through
true	thin	though	tin

twins	*tin*	three	toy
yard	too	tie	where
up	tree	use	who
us	way	we	why
wet	well	west	will
what	with	when	wood
wire	your	wool	yes
you	young	year	yet
	Removed words		
mew	pew	owes	darn
felt	star	shove	ought
bathe	oak	lie	bread
ache	tare		art
	bin		nuts
	ail		
	dumb		