TITLE
Sentence recognition in noise: variables in compilation and interpretation of tests

AUTHORS
Marianne Theunissen[1, 2]
DeWet Swanepoel[1]
Johan Hanekom[2]

AFFILIATION
[1]Department of Communication Pathology, University of Pretoria
[2]Department of Electrical, Electronic and Computer Engineering, University of Pretoria

KEY WORDS
Speech audiometry
Sentence recognition
Speech recognition in noise
Hearing in noise
Speech perception

ABBREVIATIONS

SRN:	Sentence recognition in noise
HINT:	Hearing in noise test
SNR:	Signal-to-noise ratio
LTASS:	Long-term average speech spectrum
ICRA:	International Collegium for Rehabilitative Audiology
BKB-SIN:	Bamford-Kowal-Bench Speech-in-Noise test
QuickSIN:	Quick Speech-in-Noise test
WIN:	Words-In-Noise test

CORRESPONDING AUTHOR
De Wet Swanepoel
Department of Communication Pathology
University of Pretoria
Lynnwood Road
Pretoria
South Africa
0002

E-mail: dewet.swanepoel@up.ac.za

ABSTRACT

Tests of sentence recognition in noise constitute an essential tool for the assessment of auditory abilities that are representative of everyday listening experiences. A number of recent articles have reported on the development of such tests, documenting different approaches and methods. However, both the development and interpretation of these tests require careful consideration of many variables. This article reviews and categorizes the stimulus, presentation, subject, response, and performance variables influencing the development and interpretation of tests of sentence recognition in noise. A systematic framework is utilized to document published findings on these variables. Recommendations and guidelines, based on test performance requirements and test objectives, are provided concerning the interpretation of results and the development of new test materials.

INTRODUCTION

The ability to understand speech in the presence of background noise constitutes a great challenge to any listener, especially listeners suffering from hearing impairments. Because of the challenge that this task poses to listeners, its assessment can provide great insight into an individual's ability to cope with typical everyday listening environments, as these situations are often noisy. The development of methods to assess and predict this ability has therefore received ample attention in research over several decades. The Articulation Index (French & Steinberg, 1947), and more recently the Speech Intelligibility Index (American National Standards Institute, 1997), are among measures that can provide predictions of the intelligibility of a signal in both quiet and noisy environments. Although this method of prediction has long been validated and shown to be accurate in predicting speech recognition scores in normal-hearing listeners and those with mild to moderate hearing impairment (Fletcher & Galt, 1950; Kryter, 1962; Kamm et

al, 1985), a number of researchers have demonstrated that these predictive indices have significant shortcomings in accurately predicting speech recognition scores in some populations and/or listening conditions (Humes et al, 1986; Hargus & Gordon-Salant, 1995; Ching et al, 1998; Moore, 2002). The Articulation Index and Speech Intelligibility Index provide an indication of audibility of certain speech cues, and are not direct measures of speech intelligibility (Hornsby, 2004). Other factors that have been shown to influence speech recognition, such as language abilities (Weiss & Dempsey, 2008), cognition (Humes, 2002), and spectral and temporal processing (Houtgast & Festen, 2008) are not accounted for when using the Articulation or Speech Intelligibility Index. Since these factors could affect an individual's ability to handle daily speech recognition tasks, measures that do not account for these functions will give a limited representation of everyday functioning. The limitations of speech recognition prediction measures underscore the need for direct measurement of speech recognition abilities. Additionally, many patients view their difficulty to understand speech in noise as their most important hearing problem, and consider a direct test of their ability to understand speech in noise to be of great value (Killion, 2002).

Plomp (1978) developed a quantitative model to explain the communication handicap caused by a hearing loss, and the limited benefit of hearing aids. Within this model, the loss of hearing for speech, or speech-hearing loss, is quantified according to the 50% level of speech recognition, also called the speech reception threshold. This work by Plomp indicated the need for the development of a reliable test that could be used to determine the speech reception threshold for sentences, and such a test was developed shortly after the publication of the model (Plomp & Mimpen, 1979a). The developed test was subsequently used to demonstrate the accuracy of Plomp's model (Plomp & Duquesnoy, 1982), and to quantify the effect of noise, age,

presbyacusis, hearing aids, interaural time delays and other factors on sentence recognition in noise (e.g. Plomp & Mimpen 1979b; Duquesnoy & Plomp, 1980; Festen & Plomp, 1986; Bronkhorst & Plomp, 1988).

In addition, the test developed by Plomp and Mimpen (1979a) provided groundwork that many subsequent researchers have referred to in the development of similar tests (e.g. Nilsson et al, 1994; Wong & Soli, 2005; Vaillancourt et al, 2005). Developing these tests requires careful attention to a number of variables that influence the procedures and results of the test. This article provides a systematic consideration of the variables involved in a test of sentence recognition in noise (SRN), with a two-fold purpose. Firstly, these variables are essential determinants during the development of a test for SRN, to ensure adequate test performance is attained in terms of validity, reliability, sensitivity and specificity. Secondly, these variables will influence the results of an existing test and therefore necessitate consideration for accurate interpretation of results. Both of these purposes will be addressed throughout this review, with some variables more relevant during test development, whereas others are more important in clinical application and interpretation of the test. It should be noted here that there are many existing tests of SRN with comprehensive normative data available, such as the Hearing in Noise Test or HINT (Nilsson et al, 1994) and many of its adapted versions in other languages (as listed in Soli & Wong, 2008), which provide a valuable resource for clinicians in the interpretation of test results. However, an understanding of variables that affect sentence recognition in noise remain useful in guiding the development of new tests, and in the identification of factors that could explain results that deviate from the documented norms. Table 1 summarizes the categories of variables included in this review.

Table 1: Categories of variables influencing tests of SRN

STIMULUS VARIABLES

The stimulus used in a test of SRN should receive careful consideration, as it directly affects the nature and difficulty of the test. There are three main stimulus variables that will be discussed in this section, namely the sentence material, the type of background noise used, and the speaker used for the recordings of the speech material.

Sentence material

Since 1947 researchers worldwide have developed a large variety of speech perception tests using sentence material as stimuli (Lucks Mendel and Danhauer, 1997). Among widely used tests utilizing sentence materials, there is great variation in the style and content of the sentences. The Central Institute for the Deaf (CID) Everyday Sentences (Silverman and Hirsch, 1955), for example, is comprised of a list of commonly used sentences. The Synthetic Sentence Identification (Speaks and Jerger, 1965), on the other hand, is closed-set test consisting of 10 nonsense sentences constructed to approximate the syntactic structure of English, and in such a manner that each group of three consecutive words in the sentence is meaningful, but the entire sentence is not. These sentences are frequently presented in the presence of a background noise to reduce the simplicity of the listening task, as the small closed set could make the test too easy. The sentences of the Speech Perception in Noise Test (Kalikow et al, 1977) are formulated in such a way that the predictability of the last word of each sentence (considered to be the test item) is controlled to have either high or low predictability given the sentence context. This test uses a speech babble-type noise as background noise. The Hearing In Noise

Test or HINT (Nilsson et al, 1994) is comprised of 250 meaningful, recorded sentences divided into 25 phonemically matched lists of 10 sentences each, and uses background noise matched to the long-term average of the speech material's frequency spectrum.

Style and content of sentence material

Accurate speech perception requires not only integration of acoustic cues from the speech signal, but also contextual cues, such as word familiarity, sentence complexity and word frequency. These cues are especially important when the speech signal is degraded by background noise (Needleman, 1998). The more contextual cues there are in the speech, the less reliant the listener has to be on the exact acoustic properties of the sound signal (Kalikow et al, 1977). For this reason, the style and content of sentence material used in a test of SRN is an important factor that will affect performance.

The predictability of the sentence content has been shown to influence performance, as keywords in sentences providing minimal contextual cues are more difficult to recognize than keywords embedded in a sentence with many syntactic, semantic, and prosodic cues (Hutcherson et al, 1979). The Speech Perception In Noise (SPIN) test (Kalikow et al, 1977) is based on this premise, and experimentation during its development showed that in normal-hearing listeners, the average difference in percentage of words identified correctly between high and low predictability sentences across eight lists, is 47.4%.

Other characteristics of the vocabulary used in sentence materials could also influence the level of difficulty of sentences. According the Neighborhood Activation

Model (Luce & Pisoni, 1998), a spoken word activates several lexical items in a person's memory, and word identification requires discriminating between the activated lexical items. The number of activated items, the acoustic-phonetic similarity between these items, as well as the frequency of occurrence of these items all influence discrimination. In other words, the frequency of occurrence of a word, the number of phonemically similar words or neighbors (i.e. the density of the "neighborhood"), and the frequency of occurrence of similar sounding words, affect the speed and accuracy with which perceptual decisions about these words are made (Luce & Pisoni, 1998). In normal-hearing listeners, high-frequency words are responded to 7.39% more accurately than low-frequency words; responses to words in high-density neigborhoods were 3.38% more accurate than responses to words from low-density neighborhoods, and words from low-frequency neighborhoods elicited responses 1.39% more accurate than words from high-frequency neighborhoods (Luce & Pisoni, 1998). Hearing-impaired listeners have been shown to require an increase in intensity of between 3.1 and 5.8 dB to discern the same percentage of lexically "hard" words (words with low word frequency and high neigborhood density and frequency) than lexically "easy" words (with high word frequency counts and low neighborhood density and frequency) (Dirks et al, 2001).

The implication of the Neighborhood Activation Model for tests of sentence recognition in noise is that the lexical difficulty of keywords used in sentence materials will also impact the difficulty of the SRN test. The position of keywords or target words in the sentence can also affect intelligibility, with target words occurring at the end of a sentence being 5-10% less intelligible than those occurring earlier (Bell & Wilson, 2001). If only keywords in a sentence are scored, it is important to ensure that the position of the keywords is similar across sentences, as sentences with keywords in the final position will obtain poorer scores.

Plomp and Mimpen (1979a) developed an accurate test of speech reception thresholds using sentence material in the presence of background noise. In order to compile speech material comparable with everyday speech, these researchers decided to use sentences that represent conversational speech, and did not contain proverbs, exclamations, questions, or proper nouns. Many other developers of SRN tests have developed material that is representative of everyday or conversational speech (Versfeld et al, 2000; Wong and Soli, 2005; Hällgren et al, 2006; van Wieringen and Wouters, 2008; Wong et al, 2007). A number of studies have stipulated, in accordance with the criteria of Plomp and Mimpen (1979a), that the sentences should not contain proverbs, exclamations, questions, or proper nouns (Versfeld et al, 2000; Vaillancourt et al, 2005; van Wieringen and Wouters, 2008). Some studies have added to these criteria that sentences should be syntactically complete, or at least contain a verb and a noun (Kollmeier & Wesselkamp 1997; Versfeld et al, 2000; van Wieringen & Wouters, 2008). To ensure that the style and content of sentence materials represent everyday speech, and that the material is considered acceptable by the general population, sentence material is typically rated for naturalness by native speakers (Soli & Wong, 2008). Rating is usually done on a scale of one (artificial) to seven (natural), and any sentence receiving a mean rating lower than six is revised and submitted to a second round of rating.

Homogeneous intelligibility in noise

The material used for a SRN test should be assessed for homogeneity to ensure that the sentences are more or less of equal, known difficulty. The homogeneity of the material is important to ensure the validity of the test, as test material that is not equivalent in difficulty will yield inconsistent results, and will therefore not be able to accurately measure changes in a listener's abilities. If sentence materials were to be used to assess speech recognition in the presence of background noise, its

homogeneity would have to be evaluated in the presence of the same noise before applying it to clinical populations. This is because material that is of equivalent intelligibility in quiet may not be equivalent in noise, as shown by Stockley and Green (2000). These researchers applied the Northwestern University Auditory Test No. 6 (NU-6) lists to both normal-hearing and hearing-impaired listeners in quiet and in noise and found that these lists were equivalent in difficulty when applied to both groups of listeners in a quiet condition, but when presented in noise, the lists were no longer equivalent in either group of listeners. In both groups there were no significant differences between scores on each individual list when presented in quiet, but with added noise some lists became significantly more difficult than others to both normal-hearing and hearing-impaired listeners.

In order to equalize the chance of correct recognition, sentence materials developed and recorded for a test of SRN should be submitted to a procedure for selecting sentences that are equally difficult to understand in the presence of a specific level and type of noise (Plomp & Mimpen, 1979a). This can be achieved by two distinct methods. After measuring the intelligibility thereof in normal-hearing listeners at different signal-to-noise ratios (SNRs), the mean-squared amplitude of sentences with an intelligibility poorer than average could be increased to compensate for their difficulty, and the intensity of sentences with an intelligibility higher than average could be decreased (Nilsson et al, 1994). Alternatively, sentences that deviate significantly from the average intelligibility could be rejected from the collection (Versfeld et al, 2000; Vaillancourt et al, 2005; van Wieringen & Wouters, 2008). It is also possible to combine these two methods by rejecting sentences that fall outside of a pre-determined performance range, and adjusting the intensity of the remaining sentences according to intelligibility (Wong & Soli, 2005; Wong et al., 2007).

Type of noise

The spectral and temporal properties of the speech signal and the concurrent background noise affect the results of a test of speech recognition in noise (Dreschler et al, 2001). With the variation of the SNR during the test, the masking effects of the noise depend on the relationship of its spectrum to the speech signal used (Soli & Wong, 2008). There are different types of noise reported to be efficient maskers of the speech signal in SRN tests, and many tests use noise specifically developed for their test. Whereas some tests use multi-talker babble or other speech material as background noise (e.g. Kalikow et al, 1977; Cameron & Dillon, 2007a), others use noise with a spectrum equal to the long-term average spectrum of their recorded speech material (e.g. Plomp & Mimpen, 1979a; Nilsson et al, 1994). This latter type of speech-shaped noise is used to yield high accuracy of threshold determination, and ensures that the SNR is approximately equal at all frequencies by eliminating accidental differences (such as a gender difference) between the spectrum of the speech and the noise (Plomp & Mimpen, 1979a; Nilsson et al, 1994; Wagener & Brand, 2005).

Wilson et al (2007a) investigated the difference between multi-talker babble and speech-spectrum noise as maskers for the Words-In-Noise Test (WIN). These researchers found that the majority (88%) of the normal-hearing listeners in their study performed better in multi-talker babble than in speech-spectrum noise, requiring about 2 dB better SNR to attain a 50% score in the speech-spectrum noise than in the babble noise. In listeners with hearing loss, the difference was smaller – only about a 0.7 dB difference between the two noises, with the multi-talker babble being the easier condition for 56% of the hearing-impaired listeners. The difference was ascribed to the amplitude modulations of the multi-talker babble, which led to brief improvements in the SNR. Both types of noise clearly distinguished between

normal-hearing and hearing-impaired listeners (an indication of the sensitivity of both noise types), 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 type of noise. Although the findings for hearing-impaired listeners were essentially the same for the two types of noise, multi-talker babble was finally concluded to be a more appropriate masker due to its face validity in representing everyday listening situations (Wilson et al, 2007a). In contrast, the advantage of speech-spectrum noise is its validity as a masker for sentence materials that has been employed and reported by multiple previous studies (e.g. Plomp & Mimpen, 1979a; Soli & Wong, 2008).

It is also possible to use a standardised interfering noise for different tests in different languages, provided that those languages represent the mean international long-term average speech spectrum (LTASS). This international LTASS is based on a study conducted by Byrne et al (1994), who recorded speech samples from a number of speakers for thirteen different languages. Their findings indicate that the LTASS across samples was so similar that it may be reasonable to use a universal LTASS noise for a variety of applications and languages. The International Collegium for Rehabilitative Audiology (ICRA) have developed a well-specified set of speech-like noises with spectra shaped according to gender and vocal effort which can be applied as well-specified background noise in experiments on speech recognition (Dreschler et al, 2001). This type of noise has a similar frequency spectrum to the universal LTASS (Byrne et al, 1994; Dreschler et al, 2001), and has been demonstrated to yield similar results in a speech recognition test than a noise with the same long-term frequency spectrum as the specific speech stimuli used in the test (Wagener & Brand, 2005). However, the use of a noise specifically weighted to the speech sample used in the test still appears to be a frequently selected option

reported in the literature, as shown by the large number of tests recently developed using this type of noise (Soli & Wong, 2008). The different types of noise with the advantages and disadvantages of each are summarised in Table 2.

Table 2: Advantages and disadvantages of different noise types

Another important consideration when selecting or generating noise for a test of speech recognition is the presence of fluctuations in the noise and the effect of these fluctuations on test results. Interruptions in the masking noise give listeners an opportunity to get glimpses of the speech signal and patch it together in order to recognize what was said, with slow amplitude modulations or a lower number of noise bursts per second providing a greater advantage than faster modulations (Miller & Licklider, 1950). This “masking release” effect enables listeners to take advantage of “dips” in the background noise to detect speech cues, although this ability is severely impaired in listeners with cochlear pathology or a cochlear implant (Lorenzi et al, 2006). Normal-hearing listeners could therefore perform better in a SRN task when the background noise is fluctuating as opposed to stationary, whereas some hearing-impaired subjects are not able to take advantage of these fluctuations and therefore perform similarly in stationary and fluctuating noise (Wagener & Brand, 2005; Lorenzi et al, 2006). The use of fluctuating noise could therefore improve differentiation between normal-hearing and hearing-impaired subjects, but this difference is not significant for all hearing-impaired listeners, and highly fluctuating noise has also been shown to yield larger test-retest differences and a flatter intelligibility function (intelligibility slope) than stationary noise, leading to a lower test accuracy (Wagener & Brand, 2005). However, fluctuating noise can be useful in tests of SRN when studying factors that affect masking release, such as hearing loss and age (e.g. Lorenzi et al, 2006; Dubno et al, 2002).

Speaker variables

Individual differences in vocal quality and speech production of the speaker presenting test materials could affect the results attained during speech audiometry. Wilson et al (1990) found that the sound pressure level of material recorded from a female speaker had to be increased by 11-15 dB to produce the same intelligibility scores attained with the same material presented by a 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 do, however, indicate the significance of individual differences between speakers (Wilson et al, 1990). The effect that gender or individual differences have on the intensity level of the material can be overcome by digitally adjusting these levels if material is digitally recorded (Wilson & Strouse, 1999; Nilsson et al, 1994).

According to Ostergard (1983), results attained from speakers of different genders may not compare well, especially for individuals with a high frequency hearing loss. However, in a study by Versfeld et al (2000), material presented by one of two male speakers used in a first experiment, yielded a threshold that differed only 0.2 dB from the first female speaker, compared to a difference of 1.1 dB from results 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 by 1.5 dB from the first female speaker. The analysis of variance between the results for all four speakers revealed that the speaker had a significant effect on results. These findings confirm the assertion of Wilson et al (1990) 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 (Vaillancourt et al, 2005; Cameron & Dillon, 2007a) and providing speakers with specific instructions regarding the pronunciation of the material. Instructions to the speaker should include aspects such as pronouncing the material in a natural, clear manner (Versfeld et al, 2000); maintaining clarity, pace and vocal effort (Nilsson et al, 1994); and avoiding emphasis on key words during recordings (Vaillancourt et al, 2005).

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). In addition, the speaker used for the recordings should adhere to specific criteria and follow specific instructions as listed above. The Hearing In Noise Test (Nilsson et al, 1994) and many of its adapted versions in other languages (e.g. Wong & Soli, 2005; Wong et al, 2007) minimize the possible effects of speaker variables by using pre-recorded materials, and by using the performance of normal-hearing individuals as the normative reference to identify communication handicap (Soli, 2008), allowing comparison of results across listeners, and even across different language versions of the test (Wong & Soli, 2005).

PRESENTATION VARIABLES

In addition to the content of the test material, the procedures followed during the SRN test could influence the results of the test. This relates mainly to the method used for the presentation of the material, as well as the transducer used to present the speech signal and the noise during testing. Both of these aspects are reviewed in this section.

Presentation method

Any test aimed at assessing SRN in a variety of subjects must have a way to prevent ceiling and floor effects of a test score expressed in percentage (Lutman, 1997). This means that a test scored in percentages will always have a maximum score of 100 and minimum of zero. Any test designed to assess individuals with a great range of capabilities in terms of speech perception, from listeners with slight difficulty to hear in noise to listeners with severe hearing impairments, 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. For example, if the test presents stimuli at a fixed SNR and listeners can score between 0 and 100%, listeners with a severe impairment may not be able to correctly identify any of the test items, and will thus achieve 0%. Should there be a further deterioration in the listener's abilities (due to a progressive component to the hearing loss), the test cannot indicate this, as the SNR will still be too difficult for correct recognition of any test items, and still yield the same score of 0%, even though the listener's performance has, in fact, deteriorated.

One way of overcoming this challenge is through the use of an adaptive test procedure. During such a procedure, the stimulus level of each trial is determined by

the preceding stimuli and responses (Levitt, 1970). Besides the overcoming of ceiling and floor effects, an adaptive test procedure also has the advantage of being a more efficient test method, as observations are concentrated around the region of interest (Levitt, 1970, 1978). Plomp and Mimpen (1979a) employed this method in the development of their test for sentence recognition in noise by using a fixed level of noise and adjusting 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%.

Using this method, some test administrators keep the noise level constant and adjusts the speech level adaptively (e.g. Nilsson et al, 1994), while others keep the speech stimuli at a fixed level while altering the level of the noise input according to the listener's response (Lutman & Clark, 1986). Presenting both the signal and noise monaurally under headphones, Wagener and Brand (2005) have found 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 results suggest that when presenting speech through headphones, 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. However, if the test includes spatial separation of speech and noise in the sound field by presenting the speech from a loudspeaker in front of a listener and the noise from loudspeakers to the right or left, head-shadow effects will need to be considered. Head-shadow effects are greatest at high frequencies (Moore, 1995), which could affect the audibility of these frequency components. If the noise level is altered, some of its frequency components might become inaudible at times, causing inconsistency

in the masking effects of the noise and thereby affecting results. In these cases it will be necessary to keep the noise level fixed at a level audible at both ears, while altering the speech level adaptively.

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 & Mimpen, 1979a) to 72 dB (Nilsson et al, 1994). According to Wagener (2004) 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 results depended only on the SNR, and not on the presentation level. This finding was confirmed by the findings of Wagener and Brand (2005), 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 & Brand, 2005) and does not approximate the individual's uncomfortable loudness level (Wagener, 2004).

Transducer

The presentation of test stimuli can be conducted via a number of different transducers and methods. Hällgren et al (2006), in development of the Swedish HINT, presented test stimuli through a loudspeaker positioned one meter in front of the subject. Although sound field presentation has the advantage of enabling the tester to assess listeners with hearing aids or cochlear implants, reflection of the sound from the surfaces of the enclosed test area could degrade speech intelligibility (Allen & Berkeley, 1979). These reflections may influence the speech reception thresholds, and the use of headphones is therefore preferable, especially during test development (Soli & Wong, 2008).

The original American HINT used binaural headphone presentation (Nilsson et al, 1994), whereas a number of other researchers presented stimuli monaurally (Kollmeier & Wesselkamp, 1997; Versfeld et al, 2000; van Wieringen & Wouters, 2008). Plomp and Mimpen (1979a) compared results from monaural and binaural presentation, and indicated a small difference between left ear only and right ear only presentation (0.6 dB right ear advantage). Three distinct binaural conditions were used – one with the speech signal and the noise identical at the two ears; one with identical speech signal, but the noise uncorrelated at the two ears; and one where the speech signal was identical at the two ears, but the noise only partly correlated as it would be in a diffuse sound field. The binaural condition where speech and noise was identical at the two ears yielded a 50% recognition threshold at a SNR of –7.3 dB (1.1 dB better than right monaural, and 1.7 dB better than left monaural), whereas the second condition with uncorrelated noise yielded the same performance level at a SNR of –9.6 dB, and the third condition (diffuse noise presentation) required a SNR of –8.0 dB for 50% recognition. Each of the binaural conditions therefore yielded better performance than the monaural conditions, especially the conditions where the noise was uncorrelated or only partly correlated at the two ears.

The advantage in intelligibility gained from presenting sound binaurally as opposed to monaurally, especially when phase differences between the speech signal and the interfering noise are introduced, has been known for many years (e.g. Licklider, 1948; Hirsh, 1948). In tests of speech recognition in noise, the binaural release from masking effect can be assessed using either headphones, or loudspeakers in the sound field, as transducer (Soli & Wong, 2008). When using headphones, spatial separation between the speech signal and the interfering noise can be accomplished

using head-related transfer functions as measured on a KEMAR manikin[1] (Soli & Wong, 2008). 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 can be created. Presenting the test material in this manner results in intelligibility that is nearly identical in the noise left and noise right conditions, but much lower in the noise front condition (with the noise coming from the same direction as the speech). This has been found across a number of different languages that used the same method in developing and presenting their test material (Soli & Wong, 2008), suggesting that spatial release from masking is a property of the binaural auditory system that is language independent. Due to its successful application and relatively similar results across languages, this presentation method holds significant promise for simulating everyday listening experiences where the speech and noise sources are spatially separated without the interfering effects that occur in sound field testing. Unfortunately, this method cannot be directly applied to listeners with amplification devices. These individuals need to be tested in the sound field, and site-specific norms need to be established by measuring the frequency response as well as the room effects of each loudspeaker, and using digital filters to pre-equalize the outputs of the loudspeakers. In addition, a number of normal-hearing listeners may need to be tested in the specific acoustic environment in question to enable testers to adjust the sound field norms to compensate for acoustic effects (Soli & Wong, 2008).

SUBJECT VARIABLES

The ability of a listener to understand or recognize speech stimuli in the presence of a background noise is influenced by a complex combination of factors, both in terms

[1] 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.

of the test procedure, and internal to the listener. This section will focus on the internal subject characteristics that affect speech recognition, especially in the presence of noise. The five subject characteristics considered in this discussion are hearing loss, auditory processing, age, language, and cognition.

Hearing loss

The effects that a hearing loss has on the understanding of speech can be divided into two distinct categories. The first can be described as attenuation of the sound – a threshold shift that can be compensated for by increasing the level of sounds entering the ear (Plomp, 1978). This effect has also been called the audibility component of the hearing loss, is mostly linear and predictable, and can be quantified with sensitivity measures such as pure-tone thresholds (Wilson & McArdle, 2005). This attenuation effect of the hearing loss reduces the levels of both speech signal and noise as perceived by the hearing-impaired person, but does not affect the SNR required to understand speech in the presence of noise (Plomp, 1978).

The second category of hearing loss effects is the distortion component, which reduces the clarity with which speech is perceived, even if it is loud enough to overcome the attenuation effect (Stephens, 1976; Plomp, 1978). This component can affect speech intelligibility in quiet as well as noisy situations, although its primary manifestation is in the presence of noise (Plomp, 1978). The site of the auditory lesion is one of the factors that influence the extent of the distortion effect. Listeners with conductive hearing losses (as indicated by a difference between air and bone conduction thresholds) are usually able to attain 100% correct speech recognition if the sound level is increased, since the conductive loss merely attenuates the signal (Hood & Poole, 1971; Stephens, 1976). Hearing impairments caused by cochlear or neural lesions (as indicated by an elevation of both air conduction and bone

conduction thresholds)) not only attenuates the perceived signal, but also leads to distortion of many different aspects of auditory discrimination, such as the frequency processing capacity of the cochlea, intensity coding, temporal coding, and aspects of binaural processing (Stephens, 1976). It is beyond the scope of this article to provide a comprehensive review of each of these auditory discrimination skills and their relation to SRN. What has been well-documented in the literature is that both frequency resolution and temporal resolution affect speech recognition in noise (Dreschler & Plomp, 1985; Festen & Plomp, 1983; Crandell, 1991; Thibodeau, 1991), while loudness perception as measured according to intensity difference limens only show weak correlations with SRN (Noordhoek et al, 2001; van Schijndel et al, 2001; Houtgast & Festen, 2008).

Auditory processing

Although hearing-impaired patients may have difficulty understanding speech due to a loss of peripheral hearing sensitivity, there are also individuals with normal peripheral hearing sensitivity (as defined by their pure-tone thresholds) who experience difficulty in processing speech signals (Middelweerd et al, 1990). Central auditory lesions tend to affect an individual's ability to understand speech, especially in difficult listening conditions (Crandell, 1991). 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 et al, 2001). Auditory processing difficulties are also known to be associated with a history of persistent otitis media with effusion (Bellis, 2003a). In addition, neurologic disease, neurosurgery, traumatic brain injury and aging could cause auditory processing disorders in adults (Bellis, 2003b). Individuals reporting any of these risk factors in their case history are therefore expected to score below average in a test of SRN and

should therefore be excluded from normative samples in the development of such a test.

Age

Age is a subject characteristic that is indirectly related to peripheral hearing. Although age itself is not an essential factor in speech perception, there are possible deficits in functions and processes related to speech perception that are associated with aging. The incidence of hearing loss increases with age. In the United States, 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 (National Institute on Deafness and Other Communication Disorders, 2007). An investigation by van Rooij and Plomp (1990) assessing speech perception in elderly listeners using a battery of tests showed that progressive high-frequency hearing loss accounts for the greatest amount of variance in test results, while reduced mental efficiency (general slowing of performance and reduced memory capacity) accounted for a smaller part of the variance.

Barrenäs and Wikström (2000) 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.

Language

Utilisation of linguistic information stored in the memory of a person makes up an important part of understanding sentences (Kalikow et al, 1977). For this reason, individuals with language deficits could have substantial difficulty with SRN tests. Populations whose restricted language abilities may affect their performance in speech audiometry include infants and young children; hearing-impaired persons with reduced verbal language skills; mentally retarded persons; and people with aphasia (McLauchlin, 1980). Non-native listeners have also been shown to have a reduced ability to recognize speech (presented in their non-native language) in the presence of noise (van Wijngaarden et al, 2002; Bradlow & Alexander, 2007; Weiss & Dempsey, 2008). The results acquired by van Wijngaarden et al (2002) indicated that non-native listeners in their study required a 1-7 dB better SNR to attain the same level of sentence intelligibility as native listeners. The age of acquisition of the second language appears to have an influence on SRN, as late bilinguals have been shown to require better SNRs for speech intelligibility and to derive less benefit from linguistic context in perceiving sentences (Mayo et al, 1997). Recent findings have also indicated that the speech recognition abilities of bilingual listeners in their first language appear to deteriorate as their exposure to their second language increases (Weiss & Dempsey, 2008). By implication, the linguistic abilities and number of languages that listeners are proficient in, as well as the amount of exposure to these languages should be considered in the interpretation of SRN results.

Cognition

A recent survey of twenty experimental studies on the relationship between SRN and cognition concluded that a link between the two has been demonstrated, but this relationship is secondary to the predictive effects of a hearing loss. Measures of general ability, such as IQ, were mostly unable to predict speech recognition,

whereas measures of working memory were particularly effective in demonstrating a correlation (Akeroyd, 2008). Humes (2002) found that verbal IQ accounted for 5.9% of the variance in speech recognition of elderly hearing-impaired listeners, whereas non-verbal IQ combined with an aging factor accounted for 9.4%. The combined influence of age and nonverbal IQ was one of the most powerful predictors of aided and unaided speech recognition, second only to hearing loss. Reduced memory capacity and general slowing of performance as associated with aging have also been shown to influence speech perception (van Rooij & Plomp, 1990).

Contextual redundancy of sentence material provides important clues for correct recognition thereof. However, some listeners may not be able to take full advantage of this redundancy due to limited cognitive and/or linguistic abilities. Young children, for instance, are less able to make use of semantic context to understand speech than young adults (Boothroyd & Nittrouer, 1988). Although the relation between cognitive tests and tests of speech perception is not very specific, Houtgast and Festen (2008) concluded from their review of studies on the effect that cognition has on speech recognition that the development of further tests to quantify the cognitive factors involved in speech and language processing may help to account for the variance in speech recognition noted across listeners.

RESPONSE VARIABLES

Besides the abilities and characteristics of the listener, the manner in which the subject in a SRN test responds could also influence the test results. In addition, the technique used to score the subject's responses could also have an affect on the outcome of the test. These two aspects are discussed in this section.

Response channel

In a test where the subject is required to identify or recognize a particular speech item, they may be asked to repeat aloud what they heard, or asked to write down their response (Lutman, 1997). 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 (e.g. sound-treated booth) could enhance transmission of the response.

The original HINT (Nilsson et al, 1994) and all subsequent adaptations of this test (as listed in Soli & Wong, 2008) requested subjects to repeat aloud what was heard and encouraged them to guess. 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, unclear handwriting, or typing errors), 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.

Scoring method

Generally, developers of SRN tests use a different scoring method during test development than in the final test format. With the development of the American English HINT (Nilsson et al, 1994), 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; Vaillancourt et al, 2005).

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). A number of African languages (the so-called Nguni languages) use this type of spelling rules (de Schryver & Prinsloo, 2004). In these languages, there may be many multi-syllabic conjunctions that could receive a more precise scoring if syllable scoring is used.

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 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). The “whole sentence” scoring method can be used for the adaptive measurements of recognition thresholds (SNR where 50% recognition is attained) in the final phase of test development (Vaillancourt et al, 2005), as well as in the final test format as used in clinical practice. When using this method, 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. 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 (Plomp & Mimpen, 1979a).

TEST PERFORMANCE VARIABLES

The standards that a test using sentence materials for determination of a speech recognition threshold should meet are exceptionally high (Plomp & Mimpen, 1979a). 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 influenced by its reliability, validity, sensitivity and specificity (Ostergard, 1983), all of which are reviewed in this section.

Reliability

In order for a measure to be valid, it must be reliable (Lucks Mendel & Danhauer, 1997). Reliability can be defined as the consistency of a test's results across a series of different observations (Ostergard, 1983). 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. In conventional pure-tone audiometry, a shift in threshold of 5 dB or more when retesting a frequency, as stipulated by standard audiometric procedures, might require retesting of more frequencies (American Speech-Language-Hearing Association, 2005). In hearing conservation terms, a "standard threshold shift" indicating possible damage of the hearing system is quantified as 10 dB or more (Occupational Safety and Health Administration, 2002). In the case of a test for SRN, it would be expected that a person's performance on the test remains stable, provided that their peripheral and central hearing remained constant. In this type of test that usually determines the SNR where 50% intelligibility is attained, the standard deviation of the SNR across sentences should be less than 1 dB in order to differentiate between different listening situations and different listeners (Brand & Kollmeier, 2002). A number of previous reports on tests of SRN have reported standard deviations of the error between repeated measures around 1 dB (e.g. Plomp & Mimpen, 1979a; Hagerman, 1982, 1984; Nilsson et al, 1994; Versfeld et al, 2000; Vaillancourt et al, 2005; Wong et al, 2007), indicating that such a degree of variability that is both acceptable and realistic.

This level of acceptable variability should therefore apply across different observations, such as a retest, or a test with a different test administrator. The dilemma of evaluating test-retest reliability 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). 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. Hällgren et al (2006) assessed 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. Cameron and Dillon (2007b) assessed the test-retest reliability of the “Listening In Spatialized Noise-Sentences Test” or LISN-S (Cameron & Dillon, 2007a) by re-testing 46 of the children that participated in the normative study after two months, and found only small differences (0.1 to 1.3 dB) between tests. The findings of both these studies 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.

In the case of different test administrators, presentation of the test via monitored live voice could diminish the reliability of testing by introducing greater variability of the stimulus (Konkle & Rintelmann, 1983). Therefore, pre-recorded stimuli are recommended as a standard procedure to reduce this variability (Ostergard, 1983). 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. Although there are, to our knowledge, no specific values for inter-tester reliability reported in existing literature, it may be reasonable to assume that the same variability reported for test-retest reliability should be acceptable in the case of inter-tester differences, i.e. an average difference of +/- 1 dB across test lists (Hällgren et al, 2006; Cameron & Dillon, 2007b).

The reliability of the test also depends on the stability of results across different forms or lists of the same test, also called inter-list reliability (Ostergard, 1983; Nilsson et al, 1994). The equivalence in difficulty between lists is of primary importance in a SRN 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 & Stockdell, 1980). However, due to the redundancy of sentence materials, stimuli are too easily recognized if repeated (Owens, 1983). 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. According to Brand and Kollmeier (2002), a standard deviation in SNR threshold (SNR where 50% intelligibility is attained) of 1 dB or less across sentences is required to ensure that the test can accurately differentiate between listeners with varying degrees of hearing abilities. These authors proposed that it is necessary to present at least 20 sentences during a single test in order to attain a reliable result

with a standard deviation less than or equal to 1 dB across sentences. The original American English HINT (Nilsson et al, 1994) consists of 25 lists of 10 sentences each, which could also be arranged into twelve 20-sentence lists (Vermiglio, 2008). Many of the adaptations of the original HINT consist of 24 lists of 10 sentences each, which can also be grouped into 12 lists containing 20 sentences each (de Otero et al, 2008; Cekic & Sennaroglu, 2008; Huarte, 2008; Lolov et al, 2008; Moon et al, 2008; Myhrum & Moen, 2008; Shiroma et al, 2008; Vaillancourt et al, 2008; Wong, 2008; Wong et al, 2008).

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; Wong & Soli, 2005; Vaillancourt et al, 2006; Hällgren et al, 2006; Wong et al, 2007) or looking at the standard deviation of the mean scores across subjects (Kollmeier & Wesselkamp, 1997; Versfeld et al, 2000). The most common method of ensuring inter-list equivalence in SRN tests is equalizing all the sentences through a process of elimination and/or adjusting the mean-squared amplitude of the sentences in noise, and subsequently arranging sentences into phonemically matched lists. Although many researchers who used this method did not provide a detailed motivation for the need of phonemically matched lists, the method has repeatedly demonstrated its success for ensuring inter-list reliability (e.g. Nilsson et al, 1994; Bevilacqua et al, 2008; Cekic & Sennaroglu, 2008).

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). If a test measuring SRN aims to provide an indication of an individual's ability to cope with the type of speech stimuli they encounter daily, the speech material should be

representative of everyday speech. The use of real sentences as speech stimulus, instead of single words or syllables, therefore increases the validity of a speech recognition test, if the goal of the test it to predict everyday functioning (Nilsson et al, 1994). In order to further increase the validity of the test material, the developers of the original HINT (Nilsson et al, 1994) and subsequent researchers adapting this test to other languages (as listed in Soli & Wong, 2008) have had the sentence material rated for naturalness by native speakers of the test language. This method increases the apparent or face validity of the SRN test.

In addition, the presence or absence of noise as part of the stimulus also exerts an influence on test 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 represent typical listening situations to some degree 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 & Brand, 2005). Multi-talker babble noise is more representative of the type of everyday noise that listeners find problematic than speech spectrum noise (Wilson et al, 2007a), but this noise type has been found to increase intra-subject variability in the results of a speech-in-noise test (Wagener & Brand, 2005).

Validity also relates to the correlation between the test's score and other measures of the same behaviour (Lucks Mendel & Danhauer, 1997). Previous developers of

sentence recognition tests do not commonly report on this aspect of the developed measures, although Wilson et al (2007b) conducted a study to compare performance of normal-hearing and hearing-impaired listeners on four commonly available speech-in-noise protocols (HINT; Bamford-Kowal-Bench Speech-In-Noise or BKB-SIN test; QuickSIN; and the Words-In-Noise or WIN test). The SNR where 50% recognition was obtained with two of the HINT lists (3.3 dB) compared well with the SNR yielded by the QuickSIN (4.3 dB) and the WIN (3.9 dB). In addition, the psychometric slopes of both the BKB-SIN (11.9 %/dB) and the QuickSIN (10.8 %/dB) were similar to the slope reported by Soli and Wong (2008) for the American English HINT (10.6 %/dB). However, the accuracy of the findings pertaining to the HINT test in this study is confounded by the fact that only two of the 25 lists were used, and the speech was calibrated at a level 3 dB higher than the noise level. Despite this limitation, the researchers were at least able to provide some comparison between existing tests, because different standardised tests of the same ability existed in the test language (American English). However, for tests in languages where no other established tests of the same behaviour exist, it is not possible to evaluate this aspect of validity, and it may be necessary to verify that the data collected by the test correlate with the theoretical constructs underlying it, a concept called construct validity (Ostergard, 1983). In the case of a test measuring SRN, 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 & Robinson, 1997). Due to this effect, listeners with similar audiograms may have vastly different abilities to understand speech in noise (Killion & Niquette, 2000). 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 (2008). These researchers applied their developed sentence and numbers tests to a group of cochlear implant users, and found the material a valid and feasible method of assessing speech recognition in this population. To limit the number of variables introduced in experimentation when using hearing-impaired subjects to validate test material (such as different ages, degrees and types of hearing loss, audiogram configurations and supra-threshold deficits), researchers could also simulate a hearing loss in the same normal-hearing group of subjects already partaking in a study. Past researchers have followed this method to test hypotheses by simulating certain characteristics of a hearing loss (Stuart et al, 1995; Scott et al, 2001). In the development of a SRN 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.

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). 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). Therefore, a test with 100% specificity will not falsely identify any unaffected (healthy) individuals as having the disorder that was tested for. Sensitivity and specificity influence each other in a reciprocal manner in that an increase in sensitivity usually leads to poorer specificity, which affects the overall efficiency of a test (Ostergard, 1983).

The psychometric slope of a test for SRN provides an indication of its sensitivity. A steeper slope would mean that a small change in SNR would yield a large change in performance, thereby providing a sensitive measure of changes in a listener's ability to understand speech in the presence of noise. The slope of the material developed by Plomp and Mimpen (1979a) was reported to be around 15-20 %/dB. The normative data of the American English HINT reports a 10.6 %/dB slope, whereas the average slope across 13 languages in which the HINT was adapted is reported to be 10.3 %/dB (Soli & Wong, 2008).

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 et al (2007b) 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 et al, 2007b). They also found that the BKB-SIN and HINT materials were easier and yielded higher scores in both groups of subjects (Wilson et al, 2007b). 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 have greater specificity 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 unnecessary referrals.

DISCUSSION AND RECOMMENDATIONS

The literature reviewed in this article demonstrates the numerous variables that influence tests of SRN, and should receive careful consideration during the development of such tests. For some of these variables, a specific method must be followed in order to ensure adequate test performance in terms of validity, reliability, sensitivity, and specificity. For example, the sentences used in the test should be of homogeneous intelligibility in noise. Although this homogeneity is mandatory to ensure test reliability, test developers are free to choose the method used to obtain it, as different effective methods have been documented in previous articles. It is also essential to arrange sentences into equivalent lists once the sentence collection has been finalised, as this will ensure that repeated testing of the same listener with different lists will accurately reflect changes in the listener's abilities, without the results being affected by disparities in the difficulty of different lists. Furthermore, the use of pre-recorded sentence materials instead of monitored live voice presentation is highly recommended, as individual differences between speakers will affect reliability. The recommended presentation method for tests of SRN is an adaptive method, whereby the SNR is altered and the test result is expressed in a SNR required for 50% accurate recognition. This method prevents the floor and ceiling effects that a constant stimulus method (expressing results in a percentage between 0 and 100) will have. The presentation level used during the test is not dictated by previous reports, with the only prerequisite being that both the speech and noise must be audible to the listener. As far as subject variables go, all the different aspects discussed in this review (hearing loss, auditory processing, age, language, and cognition) will influence test results. When a new test of SRN is being developed, it is therefore of critical importance that subjects participating in experiments during test development be selected carefully, with consideration given to each of these aspects. Once the test has been developed and is being applied clinically, these factors can assist test administrators in interpreting results.

Although the above-listed test variables necessitate specific choices in order to ensure sufficient test performance, there are other variables that affect test results, but all the different options are able to yield adequate test performance, depending on the specific purpose and target population of the test. These variables are summarised in Table 3, along with the possible influence they could have on test results. Note that test performance variables (reliability, validity, sensitivity, and specificity) are not included in the table, as there is not a selection of options available for these variables. Instead, these variables are indirectly determined by the other variables (stimulus, test method, subject, presentation, and response variables) and are therefore not outlined in the table.

Table 3: Variables influencing sentence recognition in noise test results that should be chosen according to the objectives of the test

In addition, during the development of a test of SRN, there are a number of documented methods or options which have all been validated and which should not directly influence the results of the test. Table 4 provides a summary of these options, along with references to previous reports in which these methods were documented. Reports referenced in the table provide a description of how the listed options were applied in the development of SRN tests and serve as useful resources to guide the development of new tests.

Table 4: Validated options for test development that have no direct influence on test results

CONCLUSION

The importance of tests measuring the ability to understand speech in the presence of background noise is underscored by the large number of reports on recently developed tests of this kind in a variety of languages. However, the collection of variables and subject characteristics that influence results, as well as the variety of documented methods of test compilation, make the development of such a test a complex task that requires careful consideration of numerous aspects. The systematic framework of variables influencing test results as presented in this article, provide an indication of factors that should be considered during test development and the interpretation of test results.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES

Akeroyd, M.A. 2008. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol,* 47, S53-S71.

Allen, J.B. & Berkeley, D.A. 1979. Image method for efficiently simulating small-room acoustics. *J Acoust Soc Am,* 65, 943-950.

American National Standards Institute. 1997. *Methods for Calculation of the Speech Intelligibility Index.* New York: ANSI S3.5-1997.

American Speech-Language-Hearing Association. 2005. *Guidelines for Manual Pure-Tone Threshold Audiometry* [Guidelines]. Available from www.asha.org/policy

Barrenäs, M. & Wikström, I. 2000. The Influence of Hearing and Age on Speech Recognition Scores in Audiological Patients and in the General Population. *Ear Hear,* 21, 569-577.

Bell, T.S. & Wilson, R.H. 2001. Sentence Recognition Materials Based on Frequency of Word Use and Lexical Confusability. *J Am Acad Audiol,* 12, 514-522.

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, 10-18.

Bevilacqua, M.C., Banhara, M.R., Da Costa, E.A., Vignoly, A.B. & Alvarenga, K.F. 2008. The Brazilian Portuguese Hearing in Noise Test. *Int J Audiol,* 47, 364-365.

Boothroyd, A. & Nittrouer, S. 1988. Mathematical treatment of context effects in phoneme and word recognition. *J Acoust Soc Am,* 84, 101-114

Bradlow, A.R. & Alexander, J.A. 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J Acoust Soc Am,* 121, 2339-2349.

Brand, T. & Kollmeier, B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J Acoust Soc Am,* 111, 2801-2810.

Bronkhorst, A.W. & Plomp, R. 1988. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J Acoust Soc Am,* 83, 1508-1516.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., et al. 1994. An international comparison of long-term average speech spectra. *J Acoust Soc Am,* 96, 2108-2120.

Cameron, S. & Dillon, H. 2007a. Development of the Listening in Spatialized Noise-Sentences Test (LISN-S). *Ear Hear,* 28, 196-211.

Cameron, S. & Dillon, H. 2007b. The listening in spatialized noise-sentences test (LISN-S): test-retest reliability study. *Int J Audiol,* 46, 145-153.

Carstens, W.A.M. 2003. *Norme vir Afrikaans: Enkele riglyne by die gebruik van Afrikaans* (4th ed.). Pretoria: Van Schaik Uitgewers.

Cekic, S. & Sennaroglu, G. 2008. The Turkish Hearing in Noise Test. *Int J Audiol,* 47, 366-368.

Ching, T., Dillon, H. & Byrne, D. 1998. Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification. *J Acoust Soc Am,* 103, 1128-1140.

Crandell, C.C. 1991. Individual Differences in Speech Recognition Ability: Implications for Hearing Aid Selection. *Ear Hear,* 12, Supplement, 100S-108S.

de Otero, C.B., Brik, G., Flores, L., Ortiz, S. & Abdala, C. 2008. The Latin American Spanish Hearing in Noise Test. *Int J Audiol,* 47, 362-363.

de Schryver, G. & Prinsloo, D.J. 2004. Spellcheckers for the South African languages, Part 1: The status quo and options for improvement. *S Afr J Afr Lang,* 24, 57-82.

Dirks, D.D., Takayanagi, S. & Moshfegh, A. 2001. Effects of Lexical Factors on Word Recognition Among Normal-Hearing and Hearing-Impaired Listeners. *J Am Acad Audiol,* 12, 233-244.

Dreschler, W.A. & Plomp, R. 1985. Relations between psychophysical data and speech perception for hearing-impaired subjects. II. *J Acoust Soc Am,* 78, 1261-1270.

Dreschler, W.A., Verschuure, H., Ludvigsen, C. & Westermann, S. 2001. ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment. *Audiology,* 40, 148-157.

Dubno, J.R., Horwitz, A.R. & Ahlstrom, J.B. 2002. Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing. *J Acoust Soc Am,* 111, 2897-2907.

Duquesnoy, A.J. & Plomp, R. 1980. Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis. *J Acoust Soc Am,* 68, 537-544.

Festen, J.M. & Plomp, R. 1983. Relations between auditory functions in impaired hearing. *J Acoust Soc Am,* 73, 652-662.

Festen, J.M. & Plomp, R. 1986. Speech-reception threshold in noise with one and two hearing aids. *J Acoust Soc Am,* 79, 465-471.

Fletcher, H. & Galt, R.H. 1950. The perception of speech and its relation to telephony. *J Acoust Soc Am,* 22, 89-151.

French, N.R. & Steinberg, J.C. 1947. Factors governing the intelligibility of speech sounds. *J Acoust Soc Am,* 19, 90-119.

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.

Hagerman, D. 1982. Sentences for testing speech intelligibility in noise. *Scand Audiol,* 11, 79-87.

Hagerman, D. 1984. Clinical measurements of speech reception thresholds in noise. *Scand Audiol,* 13, 57-63.

Hällgren, M., Larsby, B. & Arlinger, S. 2006. A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *Int J Audiol,* 45, 227-237.

Hargus, S.E. & Gordon-Salant, S. 1995. Accuracy of Speech Intelligibility Index Predictions for Noise-Masked Young Listeners With Normal Hearing and for Elderly Listeners With Hearing Impairment. *J Speech Hear Res,* 38, 234-243.

Hirsh, I.J. 1948. The Influence of Interaural Phase on Interaural Summation and Inhibition. *J Acoust Soc Am,* 20, 536-544.

Hood, J.D. & Poole, J.P. 1971. Speech audiometry in conductive and sensorineural hearing loss. *Sound,* 5, 30-38.

Hornsby, B.W.Y. 2004. The Speech Intelligibility Index: What is it and what's it good for? *Hear J,* 57, 10-17.

Houtgast, T. & Festen, J.M. 2008. On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise. *Int J Audiol,* 47, 287-295.

Huarte, A. 2008. The Castilian Spanish Hearing in Noise Test. *Int J Audiol,* 47, 369-370.

Humes, L.E. 2002. Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *J Acoust Soc Am,* 112, 1112-1132.

Humes, L.E., Dirks, D.D., Bell, T.S., Ahlstbom, C. & Kincaid, G.E. 1986. Application of the Articulation Index and the Speech Transmission Index to the Recognition of Speech by Normal-Hearing and Hearing-Impaired Listeners. *J Speech Hear Res,* 29, 447-462.

Hutcherson, R.W., Dirks, D.D. & Morgan, D.E. 1979. Evaluation of the speech perception in noise (SPIN) test. *Otolarnygol Head Neck Surg,* 87, 239 - 245

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. *Acoust Soc Am,* 61, 1337-1351.

Kamm, C.A., Dirks, D.D. & Bell, T.S. 1985. Speech recognition and the articulation index for normal and hearing-impaired listeners. *J Acoust Soc Am,* 77, 281-288.

Killion, M.C. 2002. New thinking on Hearing in Noise: A Generalized Articulation Index. *Semin Hear,* 23, 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, 46-53.

Kollmeier, B. & Wesselkamp, M. 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J Acoust Soc Am,* 102, 2412-2421.

Konkle, D.F. & Rintelmann, W.F. 1983. *Principles of Speech Audiometry.* Baltimore: University Park Press.

Kryter, K.D. 1962. Validation of the Articulation Index. *J Acoust Soc Am,* 34, 1698-1702.

Levitt, H. 1970. Transformed Up-Down Methods in Psychoacoustics. *J Acoust Soc Am,* 49, 467-477.

Levitt, H. 1978. Adaptive Testing in Audiology. *Scand Audiol Suppl,* 6, 241-291.

Licklider, J.C.R. 1948. The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise. *J Acoust Soc Am,* 20, 150-159.

Lolov, S.R., Raynov, A.M., Boteva, I.B. & Edrev, G.E. 2008. The Bulgarian Hearing in Noise Test. *Int J Audiol,* 47, 371-372.

Lorenzi, C., Husson, M., Ardoint, M. & Debruille, X. 2006. Speech masking release in listeners with flat hearing loss: Effects of masker fluctuation rate on identification scores and phonetic feature reception. *Int J Audiol,* 45, 487-495.

Luce, P.A. & Pisoni, D.B. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear Hear,* 19, 1-36.

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. *J Acoust Soc Am,* 80, 1030-1040.

Luts, H., Boon, E., Wable, J. & Wouters, J. 2008. FIST: A French test for speech intelligibility in noise. *Int J Audiol,* 47, 373-374.

Mayo, L.H., Florentine, M. & Buus, S. 1997. Age of second-language acquisition and perception of speech in noise. *J Speech Lang Hear Res,* 40, 686-693.

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.

Middelweerd, M.J., Festen, J.M. & Plomp, R. 1990. Difficulties with speech intelligibilities in noise in spite of a normal pure-tone audiogram. *Audiology,* 29, 1-7.

Miller, G.A. & Licklider, J.C.R. 1950. The Intelligibility of Interrupted Speech. *J Acoust Soc Am,* 22, 167-173.

Moon, S.K., Kim, S.H., Mun, H.A., Jung, H.K., Lee, J., Choung, Y. & Park, K. 2008. The Korean Hearing in Noise Test. *Int J Audiol,* 47, 375-376.

Moore, B.C.J. 1995. *Perceptual consequences of cochlear damage.* Oxford: Oxford University Press.

Moore, B.C.J. 2002. Response to “Articulation index predictions for hearing impaired listeners with and without cochlear dead regions”. *J Acoust Soc Am,* 111, 2549-2550.

Myhrum, M. & Moen, I. 2008. The Norwegian Hearing in Noise Test. *Int J Audiol,* 47, 377-378.

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

Needleman, A.R. 1998. Quantification of context effects in speech perception: influence of prosody. *Clin Linguist Phon,* 12, 305-327.

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. *Audiol,* 40, 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. *J Acoust Soc Am,* 95, 1085-1099.

Noordhoek, I.M., Houtgast, T. & Festen, J.M. 2001. Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-kHz region. *J Acoust Soc Am,* 109, 1197-1212.

Occupational Safety and Health Administration. 2002. Hearing Conservation. *OSHA 3074 2002 (Revised).* Available on www.osha.gov/Publications

Ostergard, C.A. 1983. Factors influencing validity and reliability of speech audiometry. *Semin Hear,* 4, 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.

Plomp, R. 1978. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am,* 63, 533-549.

Plomp, R. & Duquesnoy, A.J. 1982. A model for the speech-reception threshold in noise without and with a hearing aid. *Scand Audiol,* 11, 95-111.

Plomp, R. & Mimpen, A.M. 1979a. Improving the Reliability of Testing the Speech Reception Threshold for Sentences. *Audiol,* 18, 43-52.

Plomp, R. & Mimpen, A.M. 1979b. Speech-reception threshold for sentences as a function of age and noise level. *J Acoust Soc Am,* 66, 1333-1342.

Quar, T.K., Mukari, S.Z.M.S., Wahab, N.A.A., Razak, R.A., Omar, M. & Maamor, N. 2008. The Malay Hearing in Noise Test. *Int J Audiol,* 47, 379-380.

Roush, J. 2001. *Screening for hearing loss and otitis media in children.* San Diego: Singular-Thomson Publishing Group.

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.

Scott, T., Green, W.B., & Stuart, A. 2001. Interactive Effects of Low-Pass Filtering and Masking Noise on Word Recognition. *J Am Acad Audiol,* 12, 437-444.

Shiroma, M., Iwaki, T., Kubo, T. & Soli, S. 2008. The Japanese Hearing in Noise Test. *Int J Audiol,* 47, 381-382.

Silverman, S.R. & Hirsh, I.J. 1955. Problems related to the use of speech in clinical audiometry. *Ann Otol Rhinol Laryngol,* 64, 1234-1244.

Soli, S.D. 2008. Some thoughts on communication handicap and hearing impairment. *Int J Audiol,* 47, 285-286.

Soli, S.D. & Wong, L.L.N. 2008. Assessment of speech intelligibility in noise with the Hearing in Noise Test. *Int J Audiol,* 47, 356-361.

Speaks, C. & Jerger, J. 1965. Method for measurement of speech identification. *J Speech Hear Res,* 8, 185-194.

Stephens, S.D.G. 1976. The input for a damaged cochlea – a brief review. *Brit J Audiol,* 10, 97-101.

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. *J Am Acad Audiol,* 11, 91-96.

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. *Am J Otol,* 16, 658-663.

Thibodeau, L.M. 1991. Exploration of Factors Beyond Audibility That May Influence Speech Recognition. *Ear Hear,* 12, 109S-115S.

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. *Int J Audiol,* 44, 358-369.

van Rooij, J.C.G.M. & Plomp, R. 1990. Auditive and cognitive factors in speech perception by elderly listeners. II: Multivariate analyses. *J Acoust Soc Am,* 88, 2611-2624.

van Schijndel, N.H., Houtgast, T. & Festen, J.M. 2001. Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners. *J Acoust Soc Am,* 110, 529-542.

van Wieringen, A. & Wouters, J. 2008. LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and The Netherlands. *Int J Audiol,* 47, 348-355.

van Wijngaarden, S.J., Steeneken, H.J.M. & Houtgast, T. 2002. Quantifying the intelligibility of speech in noise for non-native listeners. *J Acoust Soc Am,* 111, 1906-1916.

Vermiglio, A.J. 2008. The American English Hearing in Noise Test. *Int J Audiol,* 47, 386-387.

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. *J Acoust Soc Am,* 107, 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. *Int J Audiol,* 44, 144-156.

Weiss, D. & Dempsey, J.J. 2008. Performance of Bilingual Speakers on the English and Spanish Versions of the Hearing in Noise Test (HINT). *J Am Acad Audiol,* 19, 5-17.

Wilson, R.H., Carnell, C.S. & Cleghorn, A.L. 2007a. The Words-In-Noise (WIN) Test with Multitalker Babble and Speech-Spectrum Noise Maskers. *J Am Acad Audiol,* 18, 522-529.

Wilson, R.H. & McArdle, R. 2005. Speech signals used to evaluate functional status of the auditory system. *J Rehabil Res Dev,* 42 (Suppl 2), 79-94.

Wilson, R.H., McArdle, R.A. & Smith, S.L. 2007b. An Evaluation of the BKB-SIN, HINT, QuickSIN, and WIN Materials on Listeners With Normal Hearing and Listeners With Hearing Loss. *J Speech Lang Hear Res,* 50, 844-856.

Wilson, R.H. & Strouse, A. 1999. Psychometrically Equivalent Spondaic Words Spoken by a Female Speaker. *J Speech Lang Hear Res,* 42, 1336-1346.

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. *J Speech Hear Disord,* 55, 771-778.

Wong, L.L.N. 2008. The Cantonese Hearing in Noise Test. *Int J Audiol,* 47, 388-390.

Wong, L.L.N., Liu, S. & Han, N. 2008. The Mainland Mandarin Hearing in Noise Test. *Int J Audiol,* 47, 393-395.

Wong, L.L.N & Soli, S.D. 2005. Development of the Cantonese Hearing In Noise Test. *Ear Hear,* 26, 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 Hear,* 28, Supplement, 70S-74S.

Table 1: Categories of variables influencing tests of SRN

Category	*Sub-categories*
Stimulus variables	Sentence material: - Style and content - Homogeneous intelligibility in noise
	Type of noise
	Speaker
Presentation variables	Presentation method
	Transducer
Subject variables	Hearing loss
	Auditory processing
	Age
	Language
	Cognition
Response variables	Response channel
	Scoring method
Performance variables	Reliability
	Validity
	Sensitivity and specificity

Table 2: 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 et al, 2007a; 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 et al, 2007a; Plomp & Mimpen, 1979a; Soli & Wong, 2008
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

Table 3: Variables influencing sentence recognition in noise test results that should be chosen according to the objectives of the test

	Variable	*Influence*
Stimulus	*Sentence material*	
	Vocabulary (keywords)	Influences linguistic complexity and should be chosen according to target population
	Position of keywords	Should be kept consistent across sentences
	Representation of everyday speech	Depends on purpose of test (if aimed at reflecting everyday performance, material should be representative of daily stimuli)
	Type of noise	
	Multi-talker babble or speech-spectrum	Multi-talker babble more representative of everyday noise
		Results attained with speech-spectrum noise can be easily compared to a number of existing tests
	Speaker	
	Male/Female	Individual differences between speakers influence test results. Material should be pre-recorded and the same speaker used if results are to be compared.
Presentation	*Presentation method*	
	Adaptive / Fixed	Adaptive method more flexible (no floor/ceiling effects), allowing assessment of greater range of listeners.
	Transducer	
	Monaural/Binaural headphones	Binaural yields better performance, especially with spatial separation of speech and noise.
	Loudspeaker/simulated sound-field conditions under headphones	Loudspeaker condition yields greater variability, but accommodates listeners with amplification devices.
Subject	*Hearing loss*	
	Auditory processing	
	Age	All influence test results as discussed. During test development, subjects should be selected carefully in order to control for these variables. During clinical application, these variables can assist in interpretation of results.
	Language	
	Cognition	
Response	*Scoring*	
	Whole-sentence scoring	Quickest way to score, and should be used once sentences have been arranged in lists

Word-by-word scoring	More accurate than whole-sentence scoring, should be used in early phases of test development
Syllable-by-syllable scoring	More accurate than sentence or word scoring, should be used in early phases of development, especially for languages with conjunctive spelling styles

Table 4: Validated options for test development that have no direct influence on test results

	Test aspect	*Options*	*References*
STIMULUS	Composition of speech material	1. Develop own/original material	
		Create sentences according to specific criteria	Plomp & Mimpen, 1979a; Wong et al, 2007; van Wieringen & Wouters, 2008; Quar et al, 2008; Shiroma et al, 2008
		Use vocabulary from children's books to formulate sentences	Vaillancourt et al, 2005; de Otero et al, 2008; Cekic & Sennaroglu, 2008; Moon et al, 2008; Myhrum & Moen, 2008
		Use existing corpus of commonly used words to formulate sentences	Vaillancourt et al, 2005; Bevilacqua et al, 2008; Luts et al, 2008
		2. Adapt existing material	
		Use existing collection of sentences developed for a different type of test/purpose	Nilsson et al, 1994; Kollmeier & Wesselkamp, 1997
		Select existing sentences from digital database	Versfeld et al, 2000; Lolov et al, 2008
		Translate & culturally adapt American HINT sentences	Wong & Soli, 2005; Hallgren et al, 2006; de Otero et al, 2008; Bevilacqua et al, 2008; Cekic & Sennaroglu, 2008; Myhrum & Moen, 2008; Huarte, 2008
		3. Combine original and adapted material	Vaillancourt et al, 2005; Wong and Soli, 2005; de Otero et al, 2008; Bevilacqua et al, 2008; Cekic and Sennaroglu, 2008; Myhrum and Moen, 2008
	Method used to equalise sentence difficulty	Re-scale intensity of sentences that are too hard / too easy	Plomp & Mimpen, 1979a; Nilsson et al, 1994; Hällgren et al, 2006; Wong & Soli, 2005; Wong et al, 2007
		Eliminating / excluding sentences that are too hard / too easy	Kollmeier & Wesselkamp, 1997; Versfeld et al, 2000; Vaillancourt et al, 2005; van Wieringen & Wouters, 2008
		Select subset or decide on re-scaling based on SNR-50 only	Plomp & Mimpen, 1979a; Nilsson et al, 1994; Wong & Soli, 2005; Wong et al, 2007
		Select subset or decide on re-scaling based on SNR-50 and psychometric slope	Kollmeier & Wesselkamp, 1997; Versfeld et al, 2000; Vaillancourt et al, 2005; Hällgren et al, 2006; van Wieringen & Wouters, 2008
PRESENTATION	Presentation method used during test development	Fixed presentation level in initial phases	Plomp & Mimpen, 1979a; Nilsson et al, 1994; Vaillancourt et al, 2005; Wong & Soli, 2005; van Wieringen & Wouters, 2008; Wong et al, 2007
		Fixed presentation level throughout	Kollmeier & Wesselkamp, 1997
		Adaptive presentation method once lists have been compiled	Plomp & Mimpen, 1979a; Nilsson et al, 1994; Vaillancourt et al, 2005; Wong & Soli, 2005; van Wieringen & Wouters, 2008; Wong et al, 2007
		Adaptive presentation method throughout	Versfeld et al, 2000; Hälgren et al, 2006; Cameron & Dillon, 2007a

RESPONSE	Response channel	Written / typed	Versfeld et al, 2000
		Verbal	Plomp & Mimpen, 1979a; Nilsson et al, 1994; Kollmeier & Wesselkamp, 1997; Versfeld et al, 2000; Vaillancourt et al, 2005; Wong & Soli, 2005; Hällgren et al, 2006; van Wieringen & Wouters, 2008