1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Adult normative data for the Speech in Babble (SiB) test

Chrysa Spyridakou[1], Stuart Rosen[2], Giorgos Dritsakis[3] and Doris-Eva Bamiou[3,4]

1 Royal National Throat Nose & Ear Hospital. London, UK

2 UCL Speech, Hearing & Phonetic Sciences, London, UK

3 Ear Institute, University College London, London, UK

4 NIHR University College London Hospitals Biomedical Research Centre, London, UK

G.Dritsakis@ucl.ac.uk

Abstract

Objective: The Speech in Babble (SiB) test assesses the perception of speech in noise in UK adults. Here we define the normal range of SiB scores to enable the use of the test in clinic.

Design: In each test, 25 monosyllabic words were played in background multi-talker babble. Listeners had to repeat the word they heard. An adaptive procedure was used to determine the Signal to Noise Ratio needed to reach 50% correct responses (i.e., the Speech Reception Threshold). Eight distinct equivalent lists were available.

Study sample: Sixty-nine normal-hearing adults (aged 20-57 years) with no reported listening difficulties participated in the study and completed the SiB test twice in both ears.

Results: Normative SiB scores varied from -0.8 dB to 3.7 dB suggesting that patients outside these limits should be considered as having abnormal scores. No statistically significant difference between ears and no effect of age or sex was found. There was 'fair' test-retest reliability.

Conclusion: The SiB test is a short, valid and reliable test that can be used in UK clinics, e.g. as part of a standard APD battery or evaluating the performance of hearing impaired patients.

1. Introduction

Understanding speech in noise is a major concern relevant to not only individuals with hearing loss [HL] (Andrade et al, 2015) but also to individuals with normal hearing thresholds, such as those diagnosed with Auditory Processing Disorder [APD][1] (Spyridakou et al., 2012; Bamiou et al., 2015). The world is noisy and people face demanding listening situations in their everyday life. Speech intelligibility is commonly assessed by speech-in-quiet audiometry; however, this test may not be sensitive to problems experienced in the presence of background noise (Wilson 2011; Andrade et al, 2015). Speech perception testing in noise is thus recommended by several professional organisation guidelines as a key step in assessing for APD (e.g. British Society of Audiology, BSA 2017; American Academy of Audiology, AAA 2010; American Speech and Hearing Association, ASHA 1996).

1.1.   Why do we need Speech in Noise (SiN) tests?

The importance of speech-in-noise (SiN) tests was first highlighted in 1970 by Carhart and Tillman who recommended that SiN tests should be part of the standard audiological test battery. Over the years several SiN tests have been developed but they are not part of the routine audiology test battery. Some of the reasons are that they may not be readily available especially in an appropriate language or accent, audiologists have difficulties in scoring them and applying the scores to clinical practice and counselling, and that they are time consuming (e.g. Wilson 2007).

SiN tests should be included in the assessment of patients with HL and also people with normal hearing who complain about listening difficulties in noisy environments (Wilson and McArdle, 2005). Since reported difficulties in hearing speech in noise are the most common referral reason for an APD assessment (Hind et al., 2011) and a common symptom in patients with APD (AAA, 2010), it has been proposed that SiN tests should be part of the APD behavioural test battery (BSA 2017). Additionally, these tests can facilitate

---

[1] Auditory Processing Disorder (APD) is the term used to categorise adults reporting listening difficulties, having normal pure-tone thresholds and abnormal scores in complex listening and/or or electrophysiological tests; these listening abilities are attributed to functional deficits in sound processing within the extended central auditory nervous system (Iliadou et al 2017).

the assessment of the overall integrity of the central auditory system, provide information to guide the counselling of patients and can be used as an outcome measure following auditory training and/or management of HL (Taylor 2003).

1.2. Types of SiN tests

There are a number of SiN tests in English that are in use, and they vary in many ways: adaptive vs fixed Signal to Noise Ratio (SNR), test material (e.g. words vs sentences) or type of noise (e.g. white or speech shaped noise vs. multi-talker babble).

Currently, the majority of SiN tests use sentences as target stimuli (e.g. Fullgrabe et al. 2015). Although it can be argued that these better reflect communication demands in the real listening world compared to SiN tests that use words, sentence recognition involves a lot more factors at the linguistic and cognitive level, e.g. working memory (McArdle and Wilson, 2008). On the other hand, word-recognition in noise is more dependent on auditory processing than on linguistic/cognitive skills because syntactic and higher level predictive abilities are not involved. Wilson et al. (2007) showed that adults with peripheral HL had significantly worse scores in the word-in-noise (WIN) test compared to normal hearing adults and suggested that the WIN test was more sensitive in diagnosing peripheral HL than the BKB-SIN and HINT tests that employ sentences. Also, McArdle and Wilson (2008) examined SRTs for over 500 monosyllabic words in noise for 24 young adults with normal hearing and showed that ~42% of the variance in the mean performance across words could be accounted for by phonetic features (voicing, place and manner), with only 3% attributable to any psycholinguistic variable (like familiarity or frequency).

Regarding the types of background noise, both speech-shaped steady-state noise and babble noise are commonly used and simulate real-world listening situations. The use of multi-talker babble in particular as a masker has several advantages (e.g. Wilson 2003). It involves many speakers talking at the same time, with none of the conversations being intelligible. It is the most common environmental background noise where listeners report problems. These problems are attributed to the fact that the background noise is aperiodic, speech-spectrum shaped (thus leading to maximal spectral overlap with the speech, and so much energetic and modulation masking), while there is minimal amplitude modulation of the envelope thus reducing opportunities for "glimpsing" (i.e. the perception of speech segments in silent or low-level intervals of the masker).

Another masking parameter is the application of fixed vs variable SNR in an adaptive procedure. For some adaptive tests, the noise level is fixed with the speech level adjusted on the basis of the subject's responses (Brand and Kollmeier, 2002), while in other adaptive SIN tests, the speech level is fixed and the noise level is adjusted (Gustafsson and Arlinger, 1994). A study by Wagener and Brandt (2005) found no differences between the effect of an adaptive procedure with fixed noise level and a similar adaptive procedure with fixed speech level on speech intelligibility. Another study by Wilson and McArdle (2012) found equivalent results on speech recognition irrespective of whether the level of the speech was fixed and the level of the noise varied or vice versa.

The vast majority of SiN tests (including the ones mentioned above) have been recorded in American English. However, it has been suggested that listeners should be assessed in their native language and using the type of English they listen to on an everyday basis (Dawes and Bishop 2007, Tabri et al. 2011). Tabri et al. (2011) showed that bilingual and trilingual adults with normal hearing performed equally well in the speech-in-quiet tests, but performed significantly more poorly for their second language and even more poorly for their third language compared to performers assessed in their first language. Dawes and Bishop (2007) similarly evaluated the SCAN-C speech-based test battery, which is recorded in American English, in UK children. The authors compared the results of 99 Oxfordshire school children aged 6–10 years with normative data obtained for children from the USA. Across all ages, the UK children scored significantly worse on two of the subtests: the filtered words and auditory figure-ground sections as well as on the composite score. The authors suggested that each anglophone country should record their own speech material.

In conclusion, the above reviewed studies show that a SiN test should be in the listener's native accent, consist of words as opposed to sentences and use a type of background noise that corresponds well with real-life listening situations. The only available test like this in a British accent is the Speech in Babble (SiB) test that uses real words as targets, spoken by a native British English talker and presented in multi-talker babble (Spyridakou et al., 2012; Bamiou et al., 2015). Spyridakou et al. (2012) compared the SiB test (together with an electroacoustic TEOAE suppression test) with two self-report measures of listening and loudness perception difficulties in adults diagnosed with APD and normal control adults without listening complaints (n=12). Bamiou et al. (2015) used the SiB test to validate three hearing questionnaires in adult patients with reported listening difficulties but normal

audiograms (n=58) vs. a normal control group (n=30). Both studies found statistically significant moderate to strong correlations between the SiB test and the patient-report measures of hearing difficulties. Information provided within the Bamiou et al. (2015) study in terms of mean and standard deviation SiB performace also showed that the test separates well patients with APD from normal adults without listening complaints (p<0.001 for both ears). However, these two studes did not provide any information on the initial stages of rigorous validation that the test underwent while their normal control groups were small, and they did not provide test-retest repeatability data, nor information on how sex and age influenced test results.

2. Aim of Study

The aim of the present study was to establish normative scores and standardise the SiB as a test to assess speech perception in noise in adults to enable use of the test in clinic.

The study was approved by the National Research Ethics Committee of the National Hospital for Neurology and Neurosurgery, the Institute of Neurology. (Registration number 09/H0716/ 46) and the Royal Free and University College London Hospitals NHS Trusts. All recruited subjects provided their written informed consent.

3. Methods

3.1. SiB Test-Protocol

The SiB test employs 8 phonemically-balanced lists of 25 monosyllabic meaningful English words as targets, spoken by a phonetically-trained adult female speaker of Standard Southern British English, and presented in multi-talker babble. The word lists were developed through a series of pilot tests (Kuranatnam 2003). Briefly, an initial list of 300 commonly used monosyllabic words was compiled and recorded in an anechoic chamber, originally at a sampling rate of 16 kHz and hence limited in bandwidth to 8 kHz. This set of words was presented to 30 listeners at fixed SNRs of -4, 0, +4 dB, with each listener hearing each word

only once. Logistic regression was used to estimate the SNRs leading to 50% correct identification of each word, i.e. the Speech Reception Threshold (SRT). All words with SRT under -11 dB SNR or in which the minimum performance level (across SNRs) was lower than 50% were removed. In addition, all words with SRT higher than 8 dB SNR or in which the minimum performance level was not 50% or higher were also removed. Further removed were all impolite or dubious words, as well as those with psychometric functions that were too shallow. A number of words that were still either too easy or too hard for the extreme values of the SNRs used were further tested on another 10 subjects at -10 dB SNR for the easily recognised words and +8 dB for the difficult words. Again, a logistic regression was performed to calculate the SRT and to ensure that all selected words had a sufficiently steep psychometric function (as described above).

Of the original 300 words, 230 survived this initial selection procedure. From these were selected 8 lists of 25 words each which were balanced across lists as best as possible for a set of 28 phonetic and phonological features including aspects like: the number of initial and final consonants; whether the vowel nucleus was a diphthong, or tense or lax monophthong; consonantal features of place, manner and voicing. Full details of the attributes used to balance the lists, and the lists themselves, can be found in the Appendix.

The masking noise was 20-talker babble obtained from the University College Hospital/Middlesex Hospital Video Laser Disc, 1993 whose spectrum was modified and filtered to more closely approximate the long-term average speech spectrum as estimated from measurements for combined male and female voices from Byrne et al. (1994). Figure 1 shows the reasonably close correspondence in spectral content between the target words and the masker.
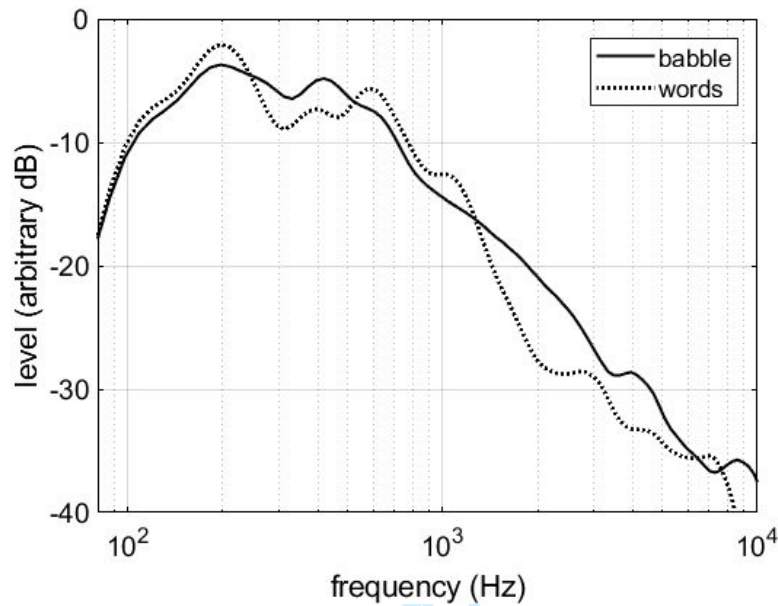
Figure 1. *A comparison of the smoothed spectra of the babble masker and the complete set of target words used in the SiB.*

The test was presented monaurally in a sound-proof room on a calibrated computer using custom-written Matlab software with Sennheiser HD 25 headphones. On each trial, a word was delivered with 500 ms of babble masker at the beginning and the end of the word itself. The listeners were required to repeat the words that they heard. Two random lists were selected (out of 8 in total) for each ear. Each ear was tested twice, to obtain information about test reliability.

The SNR was varied adaptively during the test, with the level of the masker fixed at about 65 dB SPL over the frequency range 100 Hz - 5 kHz, and the level of the target speech varied. The initial SNR was +12 dB SNR, and was decreased after each single correct response (i.e., increasing difficulty) or increased after each incorrect response (i.e., decreasing difficulty). The nominal SNR to be presented was adjusted on each trial by a word–specific correction factor in order to make the SRTs for each word equal (as determined in the preliminary test construction procedure described above). The step size changed over the course of the first few reversals from 6 dB to 2 dB SNR. The test stopped either at 8 reversals or 25 words. The SRT was then calculated as the mean of

the final six to eight reversals for a set of words for a particular listener, rather than for a particular word across listeners. Note that better performance on this measure is reflected in more negative numbers, because that implies that the criterion level of performance can be obtained at poorer SNRs. Hereafter, these SRTs will be referred to as SiB scores.

3.2. Recruitment

Adult normal-hearing volunteers were recruited from among hospital staff, hospital visitors, friends and relatives.

The inclusion criteria were as follows:

- English as first language
- Normal hearing defined as ≤20dB HL at each frequency for 250–8000 Hz at octave steps in both ears (established with pure-tone audiometry test)
- Normal middle ear function on both sides (verified with tympanometry)
- No history of psychiatric disease or cognitive impairment, as reported by the volunteers, during the initial medical interview
- No complaint of listening difficulties

Volunteers who agreed to participate in the study were given an information leaflet and signed a written consent form.

4. Results

4.1. Participants

Seventy-four potential participants were initially recruited for the study, and this included 20 subjects from the normal group reported in Bamiou et al. (2015) on whom data were available for both both first and second repeat of the SiB test in each ear. After the pure-tone audiometry tests, 5 of them were excluded since they had high-frequency, bilateral sensorineural HL (excluded subjects included two women aged 62 and 73 years and 3 men aged 67, 74 and 82 years).

Finally, 69 normal-hearing volunteers aged 20–57 years (mean age, 33.2 years; SD, 9.86) met the inclusion criteria and participated in the study. The participants included 40 women and 29 men. All participants had Jerger type A on tympanometry. They did not have any history of ear infections nor any auditory symptoms (i.e. hearing difficulties, tinnitus or hyperacusis). The mean pure-tone average for the right ear was 6.5 dBHL (SD = 3.4) and for the left ear was 6.6 dBHL (SD = 3.7). The distribution was normal for both ears, and the mean values did not differ (paired *t*-test; p = 0.71). After completion of the pure-tone audiometry and tympanometry, the participants underwent the SiB test. None of them were familiar with the test.

4.2.   SiB scores: Comparison between Ears

Table 1 summarises the mean scores and standard deviations (SDs) for the SiB test of the right and left ears for the 4 trials (2 in each ear) as well as the difference between the 2 ears and the test-retest reliability data. There were no statistically significant differences between the two ears.

Table 1: *Mean scores, standard deviations (SD), p values between the right and left ear and intra-class correlation coefficients between the first and second trial in the SiB test*

| Right ear (n = 69)		Left ear (n = 69)		*t*-test between ears for each trial
First trial mean (SD)	1.43 (±1.54)	**First trial mean (SD)**	1.26 (±2.06)	*p* = 0.54
Second trial mean (SD)	1.43 (±1.89)	**Second trial mean (SD)**	1.04 (±1.93)	*p* = 0.12
ICC (95% CI, lower-upper bound)	0.255 (0.019-0.463)	**ICC (95% CI, lower-upper bound)**	0.397 (0.180-0.578)	

4.3. Test repeatability

Test-retest reliability was assessed in 2 ways: using a paired *t* test between the 2 trials in each ear and using the 'intra-class correlation coefficient' (ICC), which measures score changes over time by comparing differences within subjects with the total variance. If individual subjects give similar responses across time, the total variance is dominated by inter-subject variability and the ICC will be high (Field 2013).

No significant differences were found between the 2 trials in either the right (p = 1) or the left ear (p = 0.4) suggesting that the test was repeatable. ICCs were calculated for each ear separately using a single-measurement, 2-way mixed effects model for absolute agreement (Koo et al. 2016). Repeatability of SiB scores across the 2 trials was judged to be 'fair' (Landis & Koch 1977) for both ears with ICCs > 0.21 (Table 1).

In order to assess repeatability among the 4 trials in both ears, a repeated measures one-factor analysis of variance (ANOVA) was used. Mauchly's test was not significant (p = .419), therefore sphericity was assumed. There were no statistical differences between the mean values of the 4 trials.

4.4. Effects of Age and Sex

Figure 2 illustrates the relationship between age, sex and the mean of the 4 SiB scores for each listener. A regression analysis was conducted to assess the effects of age and sex on the results with the SiB mean score as the outcome variable, and age and sex as predictors (Table 2).
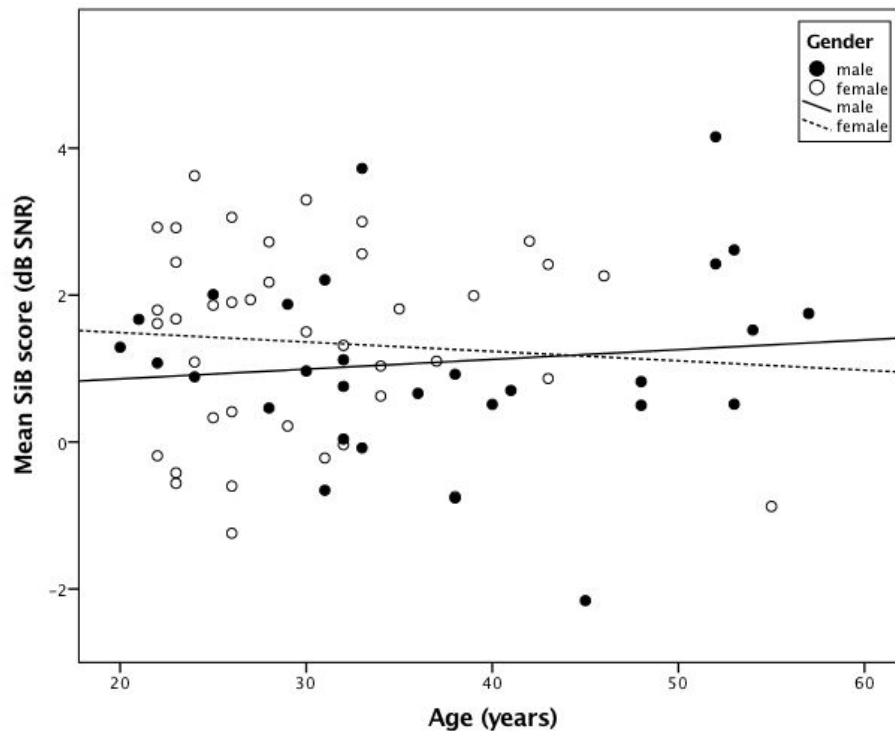
Figure 2: *SiB mean scores as a function of age shown separately for men and women.*

Both factors (age and sex) together describe only 1.1% of the variance in the SiB scores and the overall relationship was not statistically significant (p = 0.4 for sex and p = 0.9 for age. The interaction between age and sex was not statistically significant either (p=0.8). Since there is no significant effect of age or sex, these factors are ignored in the analysis of normative scores below.

4.5. Normal Range of Scores

Since there was no statistically significant difference between the values obtained in the two trials and between the two ears for the normative data, the average of all 4 measurements were used for each participant. The mean score of the SiB test in all 4 measurements was 1.3 (SD = 1.3). The data ranged from -2.2 to 4.2 and was normally distributed. Approximately 95% of the data values fell within 2 SDs from the mean and were between -0.8 dB and 3.5 dB. Therefore, 3.5 dB was taken as the upper normal range cut-off score for the SiB test in both ears. According to this, participants with scores > 3.5 dB can be considered to have abnormally poor SiB test results.

However, the above thresholds have much better precision than what a typical listener would score, as they are based on the average of 4 measurements. This is not the case in the clinic where usually one measurement is made. Therefore, the normative range was also calculated for the first trial of each subject and averaged across ears. The mean SiB score of the first measurement in both ears was 1.4 (SD = 1.4) with the scores normally distributed and ranging from -1.5 to 4.2 dB. About 95% of the data lay between -1.2 dB to 3.7 dB.

In conclusion, the upper normal limit of the SiB test is:
- 3.5 dB if 2 measurements from each ear are taken
- 3.7 dB if only one measurement from each ear is taken

In either case (as SiB scores are expressed in SRTs so higher scores indicate worse performance) patients scoring above these thresholds can be considered to have abnormal SiB results.

5. Discussion

In this study, normative scores were established for the SiB test by expanding the normal population used in previous smaller studies (Spyridakou et al. 2012, Bamiou et al. 2015) and assessing additionally age and sex effects, and test-retest repeatability. The upper cut-off for the normal range was 3.5 dB (two measurements) or 3.7 dB (one measurement) in each ear, therefore a score of > 3.5 dB or 3.7 dB, respectively, was classified as abnormal. Normative data were collected with the view to using the test in the clinic as part of the APD diagnostic test battery as well as for the assessment of other types of hearing impairment as a monaural, low-redundancy speech in noise test. Inclusion of speech based tests in the APD test battery for both neurological and non-neurological patients is deemed to be necessary, because speech stimuli are ecologically valid and tap into additional pathways to non-speech stimuli (e.g. Chermak et al. 2017). Within the context of acquired APD due to different neurological pathologies, speech based type deficits can be dissociated from non-speech type deficits and may exist in the absence of aphasia (e.g. Gutschalk et al. 2015).

There is thus a need for well-validated speech tests for the assessment of these and other patients with suspected APD.

The SiB test offers some advantages over other existing speech in noise (SiN) tests: (1) words are recorded in British English making the test more appropriate for use in the UK than other tests recorded in American English, (2) as it uses words rather than sentences, it taps predominantly into auditory rather than linguistic and cognitive skills and may therefore be more suitable in assessing auditory sensory processing problems in patients with suspected APD; in fact, it has been shown that the test correlates with patient-reported hearing difficulties and separates well patients between APD and normal adults (Spyridakou et al. 2012, Bamiou et al. 2015), (3) it focuses on one type of noise (babble noise of 20 talkers), which is a reasonable representation of real-life listening situations and it is also likely to minimise informational masking and glimpsing because of the number of talkers in the babble, (4) it is easy and quick to perform as it takes approximately 8 minutes for the completion of the four trials (two in each ear). The repeatability of the test further suggests that one trial in each ear may be sufficient, reducing the testing time even more.

No effect of ear or sex was found. This is in keeping with other studies that collected normative data for SiN tests (e.g. Wilson et al., 2007). The lack of an effect of ear may be due to the fact that in all these studies, including the present, participants had normal peripheral hearing and no auditory complaints that would indicate auditory processing problems.

Age did not affect SiB performance either. However, the upper age limit of our research participants was only 57 years, while their hearing thresholds were normal. There is mixed evidence in the literature regarding effects of age on speech perception in noise. For example, a study by Schoof and Rosen (2014) did not find any differences in sentence recognition in speech-shaped noise by normal-hearing older adults despite declines in cognitive processing compared to young ones. On the other hand, there is evidence suggesting that older adults have significantly more difficulty in hearing speech in noise than younger adults (Füllgrabe et al. 2015; Ben-David et al., 2012). For instance, Ben-David et al., 2012 showed that older adults (mean age: 72.3 years) required a significantly higher SNR for word recognition in steady-state speech-spectrum noise compared to young adults

(mean age: 20 years), and this age-related difference was higher for a babble-type masker. The authors proposed that the age-related differences were due to a combination of peripheral sensory and cognitive factors, in view of the more pronounced effects of informational masking in the older adults. However, no relative individual variability in the babble condition is mentioned that could explain the discrepancy between our findings and those by Ben-David et al. (2012). Also, Füllgrabe et al. 2015 found significantly better performance in speech perception in noise for adults 18-27 years old compared to adults 60-79 years old and suggested that this was related to cognitive decline but their target stimuli were consonants and sentences in contrast with the present study that used words.

A limitation of the current study is that normative data of individuals older than 60 years old were not collected. In order to extend use of the SiB test in clinic for older adults, further collection of normative data would be required.

6. Conclusion

Normative scores were established for a monaural British English SiB test performed on adults aged <60 years and having normal hearing and no reported listening difficulties. The test took 8 minutes to perform. No statistically significant differences were noted between ears, sexes or ages and the test had fair repeatability. In the UK, there is a paucity of validated SiN tests and the SiB is a quick, simple and reliable test with potential clinical utility for assessing speech in noise performance in adults with hearing impairments or normal-hearing adults with listening-related complaints.

Acknowledgements

London Hospitals (UCLH) Biomedical Research Centre (BRC). The funders had no involvement in the research.

References

American Academy of Audiology, 2010. "Clinical practice guidelines – diagnosis, treatment, and management of children and adults with central auditory processing"

American Speech-Language-Hearing Association (ASHA), 1996. Task Force on Central Auditory Processing Consensus Development Central auditory processing: Current status of research and implications for clinical practice. Am. J. Audiology, 5(2): 41-54.

American Speech-Language- Hearing Association (ASHA) 2005 (Central) Auditory Processing Disorders [Technical Report].

Andrade AN, Iorio MC, Gil D.2015 Speech recognition in individuals with sensorineural hearing loss. Braz J Otorhinolaryngol. Nov 6.[Epub ahead of print]

Bamiou DE, Iliadou VV, Zanchetta S, Spyridakou C. 2015 What Can We Learn about Auditory Processing from Adult Hearing Questionnaires? J Am Acad Audiol. Nov-Dec;26(10):824-37

Ben-David, BM, Tse, VY, Schneider, BA, 2012. Does it Take Older Adults Longer than Younger Adults to Perceptually Segregate a Speech Target from a Background Masker? Hear. Res. 290(1-2): 55-63.

Brand, T and Kollmeier, B, 2002. Efficient Adaptive Procedures for Threshold and Concurrent Slope Estimates for Psychophysics and Speech Intelligibility Tests. J. Acoust. Soc. Am. 111(6): 2801-2810.

British Society of Audiology, 2017. Auditory processing disorder (APD). Position Statement and Practice Guidance.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C. and Kiessling, J., 1994. An international comparison of long-term average speech spectra. *The journal of the acoustical society of America*, *96*(4), pp.2108-2120.

Carhart, R and Tillman, TW, 1970. Interaction of Competing Speech Signals with Hearing Losses. Arch. Otolaryngol. 91(3): 273-279.

Chermak GD, Bamiou DE, Vivian Iliadou V, Musiek FE. 2017. Practical guidelines to minimise language and cognitive confounds in the diagnosis of CAPD: a brief tutorial. Int J Audiol. 56(7):499-506.

Dawes, P, and Bishop, DV, 2007 The SCAN-C in Testing for Auditory Processing Disorder in a Sample of British Children. Int. J. Audiol. 46(12): 780-786.

Drennan WR, Gatehouse S, Lever C. 2003 Perceptual segregation of competing speech sounds: the role of spatial location. J. Acoust. Soc. Am. 114:2178–2189.

Dreschler, WA, 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(3): 148-157.

Field, A. (2014). Discovering statistics using IBM SPSS Statistics. 4th Ed. Sage Publications.

Füllgrabe, C., Moore, B. C., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. Frontiers in Aging Neuroscience, 6, 347.

Gustafsson, HA and Arlinger, SD, 1994. Masking of Speech by Amplitude Modulated Noise. J. Acoust. Soc. Am. 95(1): 518-529.

Gutschalk, A., Uppenkamp, S., Riedel, B., Bartsch, A., Brandt, T., & Vogt-Schaden, M. (2015). Pure word deafness with auditory object agnosia after bilateral lesion of the superior temporal sulcus. Cortex, 73, 24–35. https://doi.org/10.1016/j.cortex.2015.08.001

Hind, SE, Haines-Bazrafshan, R, Benton CL, Brassington, W, Towle, B, Moore, DR, 2011. Prevalence of Clinical Referrals having Hearing Thresholds within Normal Limits. Int. J. Audiol. 50(10): 708-716.

Iliadou VV, Ptok M, **Grech** H, Pedersen ER, Brechmann A, Deggouj N, Kiese-Himmel C, Śliwińska-Kowalska M, Nickisch A, Demanez L, Veuillet E, Thai-Van H, Sirimanna T, Callimachou M, Santarelli R, Kuske S, Barajas J, Hedjever M, Konukseven O, Veraguth D, Stokkereit Mattsson T, Martins JH, **Bamiou** DE, 2017. A European Perspective on Auditory Processing Disorder-Current Knowledge and Future Research Focus. Front Neurol. 8:622. doi: 10.3389/fneur.2017.00622. eCollection 2017.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine, 15(2), 155-163.

Kunaratnam, N, Bamiou, DE, Cohen M, 2003. Speech in Babble versus Speech in Noise in Normal Subjects and in Patients with Neurological Lesions (MSc dissertation UCL).

Landis, J.R. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics. 33. p.pp. 159–174.

McArdle, R and Wilson, RH, 2008. Predicting Word-recognition Performance in Noise by Young Listeners with Normal Hearing using Acoustic, Phonetic, and Lexical Variables. J. Am. Acad. Audiol. 19(6): 507-518.

Schoof T, Rosen S. 2015. High sentence predictability increases the fluctuating masker benefit. J Acoust Soc Am. 2015 Sep;138(3) :EL181-6

Soli, SD, Vermiglio, A, Wen, K, Laroche, C, Giguere, C, Iwaki, T, et al., 2002. Development of the Hearing in Noise Test (HINT) in New Languages. Short course presented at the 14th Convention of the American Academy of Audiology, Philadelphia, PA.

Spyridakou, C, Luxon, LM, Bamiou, DE, 2012. Patient Reported Speech in Noise Difficulties and Hyperacusis Symptoms and Correlation with Test Results. Laryngoscope 122 (7): 1609-1614.

Tabri, D, Abou Chacra, KM, Pring, T, 2011. Speech Perception in Noise by Monolingual, Bilingual and Trilingual Listeners.Int. J. Lang. Commun. Disord. 46(4): 411-422.

Taylor, B. (2003). Speech-in-noise tests: How and why to include them in your basic test battery. *The Hearing Journal*, *56*(1), 40-42.

Wagener, KC and Brandt, 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(3): 144-156

Wilson, RH, 2003. Development of a Speech- in- Multitalker Babble Paradigm to Assess Word-recognition Performance. J. Am. Acad. Audiol. 14(9): 453-470.

Wilson, RH and McArdle, RA, 2005. Speech Signals used to Evaluate Functional Status of the Auditory System. J. Rehabil. Res. Dev. 42 (4 Suppl 2): 79-94.

Wilson, RH, Mc Ardle, RA, Smith SL, 2007. An Evaluation of the BKB-SIN, HINT, Quick SIN, and WIN Materials on Listeners with Normal Hearing and Listeners with Hearing Loss. J. Speech, Lang. Hear. Res. 50(4): 844-856.

Wilson, RH, Carnell, CS, Cleghorn, AL, 2007. The Words-in-noise (WIN) Test with Multitalker Babble and Speech-spectrum Noise Maskers. J. Am. Acad. Audiol. 18(6): 522-529.

Wilson, RH, McArdle, RA, Roberts, H, 2008. A Comparison of Recognition Performances in Speech-spectrum Noise by Listeners with Normal Hearing on PB-50, CID W-22, NU-6, W-1 Spondaic Words, and Monosyllabic Digits Spoken by the Same Speaker. J. Am. Acad. Audiol. 19(6): 496-506.

Wilson , R.H. 2011 Clinical Experience with the Words-in-Noise Test on 3430 Veterans: Comparisons with Pure-Tone Thresholds and Word Recognition in Quiet J Am Acad Audiol; 22:405–423

Wilson, RH, and McArdle, R, 2012. Speech-in-noise Measures: Variable versus Fixed Speech and Noise Levels. Int. J. Audiol. 51(9): 708-712.

Appendix: List balancing

In order to balance the lists for acoustic-phonetic properties, every word was characterised as to its phonetic content from a set of 28 attributes which could be present or absent. These included:

- the number of initial (0-3) and final (0-2) consonants: 7 values
- whether the vowel nucleus was a diphthong, or tense or lax monophthong: 3 values
- the voicing status of each consonant: 2 values
- the place of articulation of each consonant: 6 values
- the manner of articulation of each consonant: 6 values
- the presence or absence of /s/: 2 values
- the presence or absence of /Σ/: 2 values

The last two sibilant attributes were incorporated because of the distinct acoustic properties of these two sounds, with relatively intense high-frequency energy that is strong in the region where speech generally has less energy, and so would make them identifiable more readily.

For example, the word *death* had the following attributes:

- 1 initial consonant, a voiced alveolar plosive (4 values)
- 1 final consonant, a voiceless dental fricative (4 values)
- Short vowel nucleus (1 value)
- No /s/ or /Σ/ (2 values)

Each word thus received a set of anywhere from 7 attributes (e.g., *hay*) to 16 (e.g., *stamp*). The balancing algorithm then determined a set of 8 lists in which every attribute appeared as equally often as possible. Generally speaking, the balance in the final 8 lists was strikingly even with differences across lists of at most 1 instance of a particular attribute.