Online Arabic and English digits-in-noise tests: Effects of test language and at-home testing

Adnan Shehabi,[1,*] Christopher J Plack,[2,3] Garreth Prendergast,[2] Kevin J Munro,[2] Michael A Stone,[2] Joseph Laycock,[4] Arwa AlJasser,[5] and Hannah Guest.[2]

[1] *Department of Audiology & Speech Therapy, Birzeit University, West Bank, Palestine*

[2] *Manchester Centre for Audiology and Deafness, School of Health Sciences, The University of Manchester, UK*

[3] *Department of Psychology, Lancaster University, UK*

[4] *Ansys UK Limited, Sheffield, UK*

[5] *Department of Rehabilitation Sciences, College of Applied Medical Sciences, King Saud University, Riyadh, Saudi Arabia*

** Corresponding author. Email: ashehabi@birzeit.edu*

Conflicts of Interest

The authors have no conflicts of interest to declare.

Abstract

Purpose

The digits-in-noise (DIN) test is used widely in research and, increasingly, remote hearing screening. The reported study aimed to provide basic evaluation data for browser-based DIN software, which allows remote testing without installation of an app. It investigated the effects of test language (Arabic versus English) and test environment (lab versus home) on DIN thresholds and test-retest reliability. It also examined the effects of test language on the correlations between DIN and audiometric thresholds.

Method

Fifty-two bilingual adults with normal hearing aged 18-35 completed Arabic and English diotic DIN tests (two sessions in the lab and two sessions at home via the web). Effects of language and environment on DIN thresholds were assessed via paired t-tests, while intra-class and Pearson's/Spearman's correlation coefficients quantified test-retest reliability and relations to audiometric thresholds.

Results

DIN thresholds were 0.74 dB higher (worse) for Arabic than English stimuli. Thresholds were 0.52 dB lower in the lab than at home, but the effect was not significant after correction for multiple comparisons. Intra-class and Pearson's correlation coefficients were too low for meaningful analysis, due to use of a normal-hearing sample with low between-subject variability in DIN and audiometric thresholds. However, exploratory analysis showed that absolute test-retest differences were low (<1.2 dB, on average) for both languages and both test environments.

Conclusions

Arabic DIN thresholds were a little higher than English thresholds for the same listeners. Employing home-based rather than lab-based testing may slightly elevate DIN thresholds, but the effect was marginal. Nonetheless, both factors should be considered when interpreting DIN data. Test-retest differences were low for both languages and environments. To support hearing screening, subsequent research in audiometrically diverse listeners is required, testing the reliability of DIN thresholds and relations to hearing loss.

Keywords

Digits-in-Noise; Arabic; Remote assessment

Introduction

The World Health Organization estimates that 430 million people suffer from disabling hearing impairment. Prevalence is expected to rise to 700 million people by 2050 (World Health Organization, 2023). Low- and middle-income nations, many of which are Arabic-speaking Middle Eastern and North African countries, are reported to exhibit disproportionately greater prevalence and burden of unaddressed hearing impairment compared to high-income countries (World Health Organization, 2018, 2023). The burden of untreated hearing loss involves greater years lived with disability, lower education and employment opportunities, cognitive and emotional disturbances, and social and economic difficulties (World Health Organization, 2018). Low socio-economic conditions, the lack of ear and hearing healthcare services, and poor public health measures are all factors associated with the greater prevalence and burden of hearing impairment in low-income countries (World Health Organization, 2013).

To address the rising prevalence and burden of hearing loss worldwide using easily accessible and cost-effective early detection measures, remote hearing screening tools have been developed and validated worldwide (for example, Jansen et al., 2010; Ozimek et al., 2009; Smits et al., 2004; Van den Borre et al., 2021; Watson et al., 2012). The digits-in-noise (DIN) test, which assesses the perception of digit triplets (e.g., 3-5-1) in background noise, is available in many world languages, including Dutch, French, German, Persian, Polish, and English (Folmer et al., 2017; Jansen et al., 2010; Ozimek et al., 2009; Smits et al., 2004, 2006; Watson et al., 2012; Wolmarans et al., 2021; Zadeh et al., 2021; Zokoll et al., 2012).

Typically, thresholds for DIN stimuli are obtained using adaptive methods, whereby the signal-to-noise ratio (SNR) is varied across a block of trials to estimate the 50%- or 71%-correct digit identification point on the psychometric function (Smits et al., 2013; Smits & Houtgast, 2007; Vlaming et al., 2014). The high precision of DIN threshold estimation is thought to stem from the steep psychometric functions found across many DIN test language variants (including Arabic), optimizing the diagnostic power of the DIN (Koifman et al., 2016; Van den Borre et al., 2021). Since digits are familiar, the DIN has the advantage of strongly reflecting peripheral auditory function, due to the minimal impact of top-down cognitive processing on test performance compared with other speech-perception-in-noise tasks, which require higher cognitive and linguistic processing of more complex speech materials (Heinrich et al., 2015; Smits et al., 2013).

The earliest remote DIN application was through landline telephones (Smits et al., 2004). Later versions were delivered using internet browsers and smartphone applications (Buschermöhle et al., 2014; Buschermöhle et al., 2015; Folmer et al., 2017; Jansen et al., 2010; Potgieter et al., 2016; Smits et al., 2006; Vlaming et al., 2014). Landline telephone networks are characterized by a markedly limited frequency bandwidth, typical around 0.3–3.4 kHz, excluding useful higher-frequency speech information (Smits et al., 2004). Normative data by Koifman et al. (2016) showed that thresholds for band-pass-filtered DIN stimuli presented through landline telephones were higher (i.e., worse) than those for broadband DIN stimuli presented through high-quality laboratory headphones. These differences were evident across DIN tests in Arabic, Hebrew, and Persian (about 2 dB difference for Arabic and Hebrew and 5 dB for Persian). Similar effects have been reported in European languages such as French, Polish, Dutch, and German, with limited-bandwidth telephone delivery elevating thresholds by up to 5 dB (Jansen et al., 2010; Ozimek et al., 2009; Smits et al., 2004; Zokoll et al., 2012).

High-fidelity laboratory headphones enable wideband stimuli and thus represent the gold standard for measuring auditory performance. However, consumer headphones and earphones with variable quality and specifications are increasingly being used to deliver auditory stimuli in remote psychoacoustic experiments (Hyvärinen et al., 2023). Auditory stimuli delivered by consumer headphones and earphones contain more speech information than those of landline telephones due to their wider frequency bandwidth, despite variable frequency responses beyond approximately 8 kHz (Hyvärinen et al., 2023). Vlaming et al. (2014) showed that DIN thresholds obtained at home using low-quality commercial headphones were within 1.5 dB of those obtained using higher-quality counterparts, in both normal-hearing and hearing-impaired listeners. It is important to highlight that other factors such as internet speed, device processing power, background noise, and environmental distractions may add to the measurement error of DIN thresholds performed remotely using the internet (Vlaming et al., 2014). Yet, there is little quantification in the literature of the effects of such factors on DIN thresholds.

The slopes of psychometric functions for DIN tasks measured with speech-shaped-noise maskers are reported to be steep: between 15% and 20%/dB (Van den Borre et al., 2021), producing fairly comparable thresholds across several languages despite differences in language characteristics (Koifman et al., 2016; Zadeh et al., 2021; Zokoll et al., 2012). Mean DIN thresholds obtained from normal-hearing listeners of Arabic, Hebrew, and Persian using broadband headphones were -9.3 dB (standard deviation [SD] = 0.7 dB), -10.1 dB (SD = 0.6 dB), and -9.6 dB (SD = 0.5 dB) respectively (Koifman et al., 2016). Smits et al. (2016) compared English and Dutch DIN thresholds in normal-hearing Dutch university students with good English proficiency. The authors found that English thresholds tended to be lower than Dutch, though only in one of two masker noises was this effect statistically significant. However, it is important to note that in this study, as in the presently reported study, the talkers differed between languages. In this circumstance, the potential combination of language effects and talker-specific effects complicates interpretation of findings.

For the DIN to serve as a useful hearing screening test, it must also be reliable. Several studies have assessed the reliability or variability of DIN thresholds in a number of languages, with generally promising results (Van den Borre et al., 2021). For instance, Jansen et al. (2010) assessed the variability of French DIN thresholds measured via telephone and headphones, in both normal-hearing and hearing-impaired listeners. The observed intra-individual SD (the root mean square of the within-subject SDs of repeated DIN threshold measurements) was consistently <1 dB, regardless of participant hearing status or mode of delivery. Vlaming et al. (2014) calculated the same metric of variability for English DIN tests, and obtained values that ranged from 0.7 to 1.3 dB, dependent on the frequency content of the stimuli and the hearing status of the listeners; minor learning effects were also evident. Dillon et al. (2016) administered repeated Australian-English DIN tests via telephone and found that the SD of test-retest differences was 1.71 dB and the inferred SD of individual scores was 1.21 dB; a non-significant trend for improvement from test to retest was also observed (mean test-retest difference = 0.34 dB, $p > .05$).

In summary, the literature would benefit from further data on how DIN thresholds obtained in the lab (using gold-standard equipment) compare to those measured remotely (in the listener's home environment, using their own device and headphones/earphones). Moreover, the effects on DIN outcomes of using Arabic versus English stimuli are unknown. The effects of each of these factors on test-retest reliability also require investigation. To address these aspects, the current study aims to determine: (a) how DIN thresholds are affected by test language (Arabic versus English) and test environment (lab versus home); (b) how the test-retest reliability of DIN thresholds compares across the two test languages and the two test environments; and (c) how the strength of the correlations between DIN and PTA thresholds compares across the two test languages. To achieve these aims, 52 young listeners with normal hearing, proficient in Arabic and English, were tested in two lab-based and two home-based sessions. In the lab sessions, standard and extended-high-frequency (EHF) PTA thresholds, as well as Arabic and English DIN thresholds, were measured. The home sessions involved Arabic DIN testing via the web using the participants' own equipment. Test and retest data were collected and compared for both languages and environments.

Method

Participants

A sample of 52 English- and Arabic-speaking young adults (25 females) aged 18 – 35 (mean age = 21.8; SD = 4.8) took part. Participants had hearing thresholds within normal limits (≤20 dB HL at 0.25, 0.5, 1, 2, 4, and 8 kHz) bilaterally with no self-report of past head or neck traumas or past/current diagnosis of otologic or cognitive/memory deficits. Normal outer and middle-ear functions were established by otoscopic examination and type-A tympanogram (compliance 0.3 to 1.6 cm^3; middle-ear pressure -50 to +50 daPa). Five participants of the 57 originally recruited were excluded from the study due to abnormal middle-ear function (n = 3) or hearing thresholds outside the clinically normal range (n = 2). Regarding language proficiency, recruitment criteria required only that participants were bilingually fluent. In practice, all self-reported as native Arabic speakers who had learned English in childhood. All were living in the United Kingdom at the time of participation and working or studying in English-speaking roles/educational programmes. All were judged to demonstrate bilingual fluency by the study's bilingually fluent researcher (assessed informally in the course of email exchanges and in-person communication at

the testing sessions). The required sample size was based on obtaining 80% power to detect a medium effect (Cohen's d of 0.5), at a one-tailed alpha error rate of 0.0083 (studywise error rate of 0.05, corrected for six multiple comparisons).

Study design

After providing written informed consent, participants attended two lab-based test sessions and also completed two sessions at home via the web. DIN tasks were completed at all four sessions; the lab sessions included additional audiological measures. The two lab sessions took place on different days within one month of each other. The first home session was completed the day following the first lab session; the second home session was completed on the day preceding the second lab session. The first and second lab sessions lasted for 60 and 20 minutes, respectively. Each of the home sessions lasted for 10-15 minutes. Upon the completion of the study, participants were compensated financially for their travel expenses and participation time. The University of Manchester Research Ethics Committee approved all study procedures (approval reference: 2021-12892-20911).

Audiometric thresholds

At the first lab session, air-conduction PTA thresholds were measured at 0.25, 0.5, 1, 2, 4, and 8 kHz, following the recommended procedures of the British Society of Audiology (British Society of Audiology, 2018). The same procedures used to measure standard PTA thresholds were used to determine EHF thresholds at 10 and 14 kHz. For each participant, standard and EHF PTA threshold averages were defined as the mean air-conduction thresholds at 0.5, 1, 2, 4, and 8 kHz and at 10 and 14 kHz, respectively, averaged across both ears. Audiometric thresholds were measured using a GSI Pello audiometer (Grason-Stadler Inc., Minnesota, USA) through Sennheiser HDA-300 supra-aural headphones (Sennheiser Electronic GmbH and Co. KG, Wedemark, Germany) in a double-walled sound-treated booth.

Tympanometry

Tympanometry was performed in the first lab session to ascertain normal middle-ear function. Middle-ear admittance and pressure were measured bilaterally using a GSI Tympstar (Grason-Stadler Inc., Minnesota, USA), following the recommended procedures of the British Society of Audiology (British Society of Audiology, 2013).

DIN testing environment

In the lab and home sessions, the DIN tasks were delivered using the same browser-based software: the Manchester Online Speech-perception Suite (MOSS). Participants listened via headphones or earphones (see environment-specific equipment details below) and entered responses via a mouse and on-screen number pad. In the two lab sessions, participants completed the English and Arabic DIN tasks (in random order) in a double-walled sound-treated booth using Sennheiser HD650 circumaural headphones. In the home sessions, they completed the Arabic DIN using their own computer and headphones/earphones via an individualised browser link and were instructed to do so in a quiet, distraction-free room. A demonstration version of the software is available at: https://mosspublic2.s3.eu-west-2.amazonaws.com/moss.html?&ID=demo&Experiment=demo. Note that the demonstration is intended only to provide an overview of the software environment and listener experience, not a fully realized DIN test with realistic parameters.

DIN presentation level

In both test environments, the overall stimulus level (rather than the target level or masker level) was held constant, in order to prevent loudness discomfort. In the lab sessions, the stimulus level was fixed at 63.2 dB SPL. In the home sessions, the stimulus level was determined by the participant in a subjective calibration stage, designed to ensure that all stimuli were both audible and comfortable. During the calibration, the participant was presented with a low-level sentence (“This sentence should be clear”) and a high-level sentence (“This sentence should be loud but not uncomfortable”), separated in level by 25 dB. They were instructed to adjust the volume control of their device until the low-level sentence was clearly audible and the high-level sentence was loud but not uncomfortably so. The root-mean-square

level of all subsequent stimuli was 5 dB below the high-level calibration sentence and 20 dB above the low-level sentence, ensuring that the digits did not become inaudible even at the lowest SNRs.

DIN stimuli

The target comprised a carrier phrase ("The digits…") followed by a random sequence of three digits sampled without replacement from the range 1-9. Targets were constructed from wideband digit recordings made at the University of Manchester (though note below that stimuli were subsequently filtered to remove EHF energy). The talkers were both middle-aged women, one speaking digits in British English, the other in Modern Standard Arabic, each in their native language. Both were naturally clear talkers with normal articulation rather than professional actors or voiceover artists, in line with International Collegium of Rehabilitative Audiology (ICRA) guidance (Akeroyd at al., 2015). Digit sound files were randomly selected from six exemplars of each digit (all spoken by the same talker, but with some natural variation in enunciation).

Note that each digit exemplar could occur in any of the three positions within a triplet, rather than being recorded and played at the same position in the sequence to achieve natural prosody. Multisyllabic digits were not excluded. In both respects, our approach does not conform to ICRA recommendations (Akeroyd et al., 2015) but is consistent with the subsequent response to those recommendations by Smits (2016). Note also that digits were not optimized to adjust for differences in intelligibility among individual digits, an omission that runs counter to ICRA guidance and likely added unwanted variability to the adaptive tracks (see the Discussion for further consideration of this limitation).

The digits were band-pass filtered between 0.12 and 8 kHz, to prevent discrepancies in the frequency responses of headphones and earphones outside of this range from influencing performance. The carrier phrase and the digits were embedded in speech-spectrum-shaped Gaussian masking noise of the same bandwidth as the target speech. A fresh noise token was selected at random from a longer noise file for each stimulus. The masking noise commenced 200 ms before the start of the carrier phrase and concluded 100 ms after the end of the final digit. Inter-digit intervals varied randomly in the range 180-250 ms. The "level" of a digit triplet (for the purposes of setting the SNR) was the level of the individual digits in that triplet, not the level of the concatenated triplet including silent gaps. The carrier, digits, and masker were presented diotically.

DIN threshold determination

All DIN tasks used the same adaptive methods. In each trial, correct entry of at least two out of three digits was considered a correct response. Visual feedback on response correctness was presented to participants following each trial. Stimulus SNR commenced at +8 dB and varied following a two-down one-up tracking rule. For the first two turnpoints, the SNR varied in 6-dB steps. For the last six turnpoints, the step size was 2 dB. The threshold was defined as the mean of the SNRs at the final six turnpoints. Each test block lasted approximately 5 minutes and mean block length was 26.5 trials (SD = 4.2). Testing was preceded by an unscored practice block containing just two turnpoints.

Statistical analyses

IBM Statistical Package for Social Sciences version 26.0 (IBM Corp, Armonk, USA) was used to analyse the data. The primary analyses aimed to determine whether there were statistically significant differences between (1) Arabic and English DIN thresholds (obtained in the first lab session), (2) lab-based and home-based Arabic DIN thresholds (obtained in the first lab session and first home session), (3) test-retest reliability of the Arabic and English DIN thresholds (obtained in the lab), (4) test-retest reliability of the lab-based and home-based Arabic DIN thresholds, (5) the strengths of the correlations with mean PTA thresholds of the English and Arabic DIN thresholds (obtained in the first lab session), and (6) the strengths of the correlations with mean EHF audiometric thresholds of the English and Arabic DIN thresholds (obtained in the first lab session).

Paired-sample t-tests were used to address aims 1 and 2. For aims 3 and 4, test-retest reliability was quantified via one-way, random-effects, single-rater intra-class correlation coefficients (ICC[1,1]; Shrout and Fleiss, 1979). For aims 5 and 6, the correlation strength between DIN thresholds and mean audiometric thresholds was quantified using Pearson's and (for non-normally distributed data) Spearman's correlation coefficients. For the primary analyses, the alpha level was adjusted for six

multiple comparisons using the Bonferroni–Holm method, with a studywise error rate <0.05. All statistical tests were two-tailed.

Pre-registered protocol

A detailed study protocol was pre-registered on the Open Science Framework prior to data collection (https://osf.io/58zkp). All pre-registered objectives, tools, procedures, and statistical analyses were applied as specified in the protocol, with two exceptions. Both exceptions were due to an important limitation of the study: recruitment of only normal-hearing participants, which greatly restricted between-subject variance in DIN and PTA thresholds. As a consequence, all Pearson's, Spearman's, and intraclass correlation coefficients were statistically non-significant or (in the case of the ICC for the home-based Arabic DIN) marginally significant.

Firstly, since formal comparison of non-significant correlation coefficients is not informative, no such comparisons were performed. However, 95% confidence intervals for the estimated correlation coefficients are presented (overlapping in all cases).

Secondly, in addition to the planned analyses, exploratory analyses were performed to further examine the sources of test-retest variability in DIN thresholds (since reliability is a poor metric in a sample with low between-subject variance). To better understand the variability of our DIN measures, we additionally computed absolute within-subject test-retest differences in DIN thresholds, for each of the languages and test environments. Care has been taken throughout this document to distinguish these from planned analyses.

Results

The effect of language on DIN thresholds: Arabic versus English stimuli

Figure 1A shows Arabic and English DIN data obtained in the first lab session. Arabic thresholds (mean = -10.75 dB, SD = 0.92 dB, 95% CI [-11.01 dB, -10.50 dB]) were higher than English thresholds (mean = -11.49 dB, SD = 1.10 dB, 95% CI [-11.75 dB, -11.49]). Although the mean difference was fairly modest (0.74 dB), a paired-samples t-test showed it to be highly statistically significant (t (51) = 4.20, p < .0001), even after correction for six multiple comparisons.

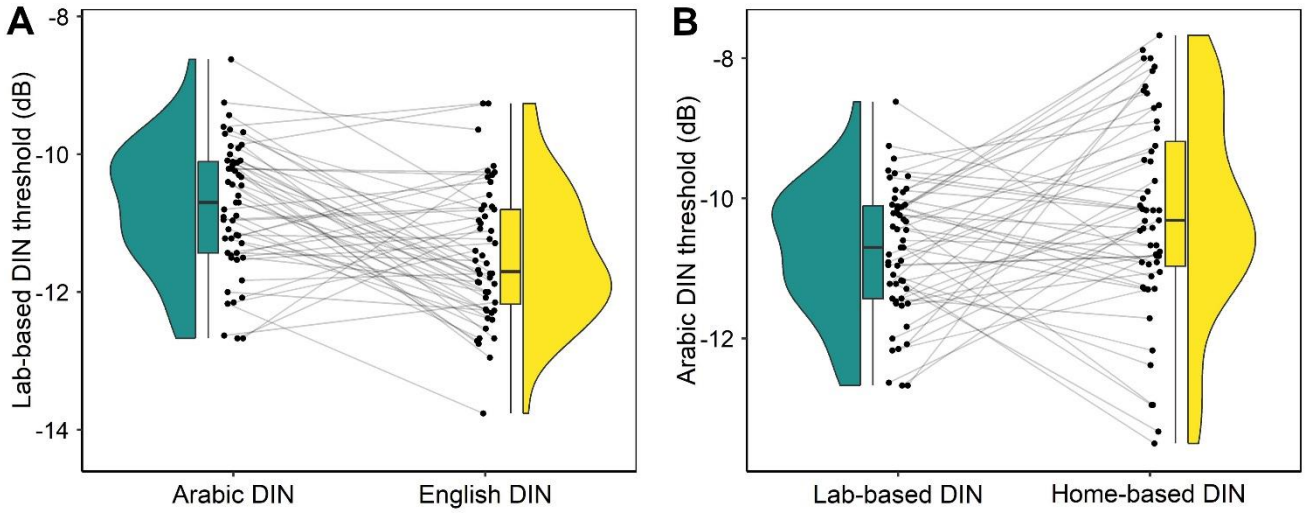

Figure 1: *Raincloud plots of the effects of test language and test environment, with marginal density plots and boxplots. The upper and lower hinges of the boxplots represent the first and third quartiles; the thick line, the median; the upper and lower whiskers, the range. A: Arabic and English DIN thresholds obtained in the first lab session. B: Arabic DIN thresholds obtained in the first lab session and first home session.*

The effect of test environment on DIN thresholds: Lab versus home

Figure 1B illustrates the relation between lab-based and home-based Arabic DIN data obtained in the first lab session and first home session, respectively. The mean of the lab-based thresholds (-10.75 dB, SD = 0.92 dB, 95% CI [-11.01 dB, -10.50 dB]) was lower than the mean of the home-based thresholds (-10.23 dB, SD = 1.44 dB, 95% CI [-10.63 dB, -9.83]). However, the difference was small (0.52 dB) and was not statistically significant after correcting for multiple comparisons ($t(51) = -2.37$, $p = .021$).

Test-retest reliability

Figure 2 shows the Session-1 and Session-2 thresholds for the lab-based Arabic DIN, lab-based English DIN, and home-based Arabic DIN. ICCs were used to quantify test-retest reliability. For the lab-based Arabic DIN, ICC (1,1) = 0.152, 95%CI [-0.122, 0.405]; for the lab-based English DIN, ICC (1,1) = 0.093, 95%CI [-0.181, 0.354]; for the home-based Arabic DIN, ICC (1,1) = 0.328, 95%CI [0.065, 0.549].

Since all ICCs were statistically non-significant or marginal (due to low between-subject variance in DIN thresholds), formal comparison of ICCs was not conducted. 95% CIs clearly overlap in all cases. Exploratory analyses showed that the mean absolute within-subject difference between the thresholds at Session 1 and Session 2 was 1.12 dB (SD = 0.68 dB) for the lab-based Arabic DIN, 1.18 dB (SD = 0.83 dB) for the lab-based English DIN, and 1.17 dB (SD = 0.92 dB) for the home-based Arabic DIN.

Correlations between DIN and audiometric thresholds

Figures 3A and 3B show the relations between DIN thresholds obtained in the first lab session and standard PTA thresholds. For the Arabic DIN, Pearson's $r(50) = 0.068$ ($p = .663$); for the English DIN, $r(50) = 0.144$ ($p = .308$). The 95% confidence intervals for the correlation coefficients, calculated using Fisher's transformation, are overlapping: [-0.209, 0.335] for the Arabic DIN and [-0.133, 0.402] for the English DIN.

Figures 4A and 4B demonstrate the relations between DIN thresholds (obtained in the first lab session) and mean EHF audiometric thresholds. The latter were not normally distributed, and hence their relations to DIN thresholds were quantified primarily via Spearman's correlation coefficients: $r_S(50) = -0.135$ ($p = .339$) for the Arabic DIN and $r_S(50) = 0.240$ ($p = .086$) for the English DIN. The 95% confidence intervals for the correlation coefficients, calculated using Fisher's transformation, are overlapping: [-0.394, 0.143] for the Arabic DIN and [-0.035, 0.481] for the English DIN.

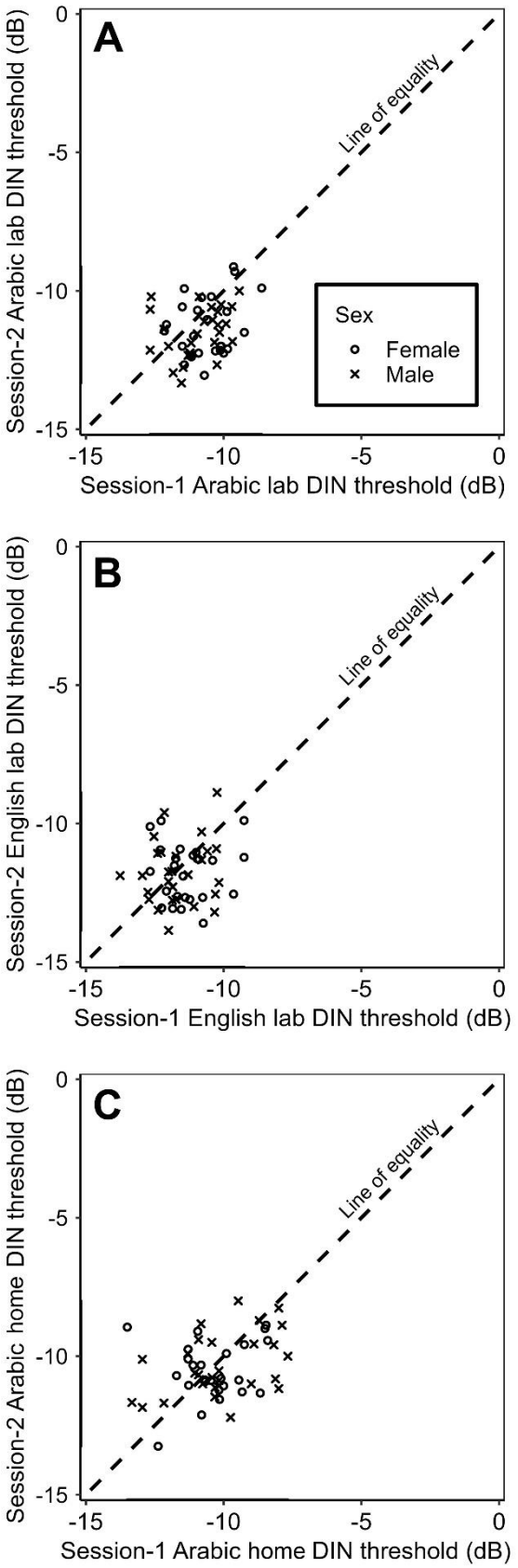

Figure 2: *DIN test-retest data. A: Arabic DIN thresholds measured in the lab. B: English DIN thresholds measured in the lab. C: Arabic DIN thresholds measured at home*

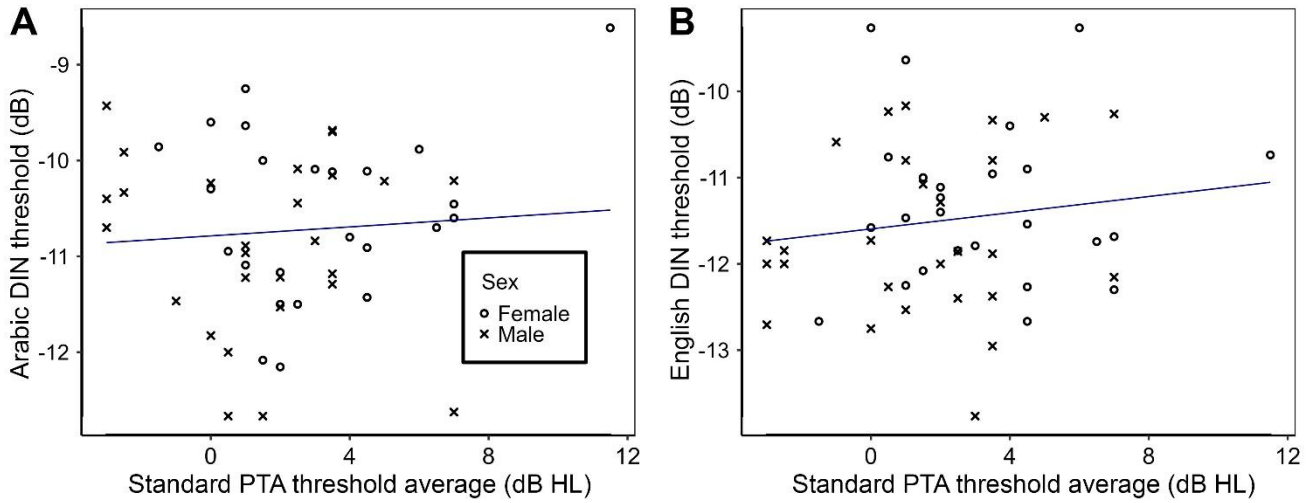

Figure 3: DIN thresholds in relation to mean PTA thresholds. Since all participants exhibited normal audiometric thresholds, no significant relations to PTA were observed. A: Arabic DIN thresholds (obtained in the first lab session) as a function of standard PTA thresholds. B: English DIN thresholds (obtained in the first lab session) as a function of standard PTA thresholds.

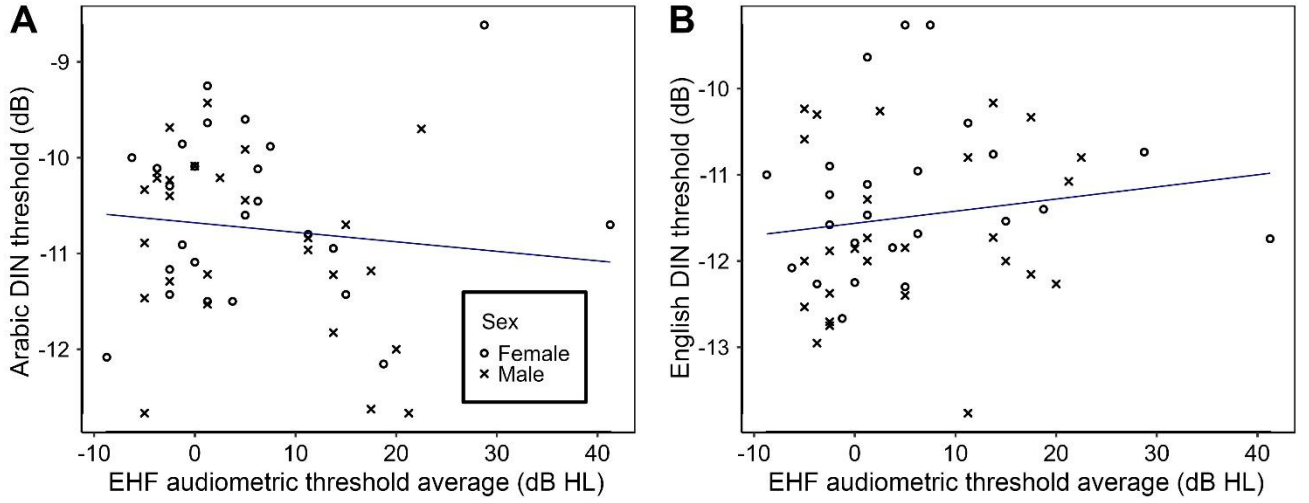

Figure 4: *DIN thresholds in relation to EHF audiometric thresholds. No significant relations were observed. A: Arabic DIN thresholds obtained in the first lab session as a function of EHF thresholds. B: English DIN thresholds obtained in the first lab session as a function of EHF thresholds.*

Discussion

The effect of language on DIN thresholds

The effect of test language on DIN thresholds was small (mean difference 0.73 dB) but significant. Despite the participants being native Arabic speakers, English DIN thresholds were lower (i.e., better) than their Arabic counterparts. This indicates that the English digits were easier to perceive, and that language proficiency is unlikely to explain the observed difference across languages. Smits et al. (2016) reported a similar pattern of findings in Dutch and English DIN with Dutch native speakers, in that English DIN thresholds were lower than their Dutch counterparts. The underlying causes of such inter-language differences are complex and beyond the scope of this publication, but it is notable that the Arabic digits have more complex syllabic structures than the mostly monosyllabic English digits.

An important caveat is that differences in intelligibility between our Arabic and English digits cannot be definitively attributed to language alone, since the talkers also differed (Hochmuth et al., 2015; Smits et al., 2016). Care was taken to ensure that the talkers shared important features, in that both were female,

middle-aged, naturally clear talkers, not actors or voiceover artists, with normal articulation. Nonetheless, one cannot rule out the possibility that the voice of our English talker – rather than English per se – was easier to perceive.

Regardless of whether the systematic difference in intelligibility between our stimuli are due to language or talker, results confirm the importance of obtaining comprehensive validation data for any newly created set of digit stimuli (Akeroyd et al., 2015). DIN screening tools generally depend on validation data comparing DIN thresholds to pure-tone audiograms in listeners with a range of hearing losses. Our data make clear that DIN validation data obtained in one language must not be used as the basis for screening tools in another, due to differences in intelligibility across languages and talkers.

The effect of test environment on DIN thresholds

Test environment had an even smaller effect, which was significant before correction for multiple comparisons but did not survive correction. Arabic DIN thresholds obtained using high-fidelity headphones in the lab were, on average, 0.52 dB better than those obtained at home using the listener's own computer and headphones/earphones. This difference is surprisingly modest, given potential deficits in the quality of each listener's headphones/earphones, computing device, internet connection, listening environment, and the absence of support from the researcher. The difference is similar to that reported by Vlaming et al. (2014) for laboratory-grade headphones vs cheap consumer headphones, though those researchers held constant the other aspects of the testing environment. The degree of departure from lab-acquired results might have been greater if participants had used a telephone or speakers (as opposed to headphones) for at-home DIN measurement (Buschermohle et al., 2014; Smits et al., 2006).

The similarity of the lab-based and home-based DIN thresholds is encouraging, suggesting that browser-based DIN testing may be feasible for remote hearing screening in the Arabic-speaking world. However, it may be necessary to apply a small correction for the test environment when comparing home-based and lab-based data.

It is also important to mention some features of our DIN testing software that may explain its good performance at home. First, our stimuli only contained energy up to 8 kHz, since consumer headphones/earphones tend to perform adequately in this frequency range but vary at higher frequencies (Hyvärinen et al., 2023). Second, the test interface was designed to provide clear and friendly instructions, demonstrations, and feedback throughout the testing process, compensating for the absence of a human tester to provide support and encouragement. Without these features, remotely measured DIN thresholds might diverge more widely from gold-standard lab-acquired data.

Test-retest reliability

Test-retest reliability as measured by ICC(1,1) was low for both languages and environments. Although this meant that we observed no significant effect of language or environment on DIN reliability, this result is not informative, because it is the consequence of a methodological limitation: recruitment of only normal-hearing participants. This homogeneity resulted in low between-subject variability in DIN thresholds. By definition, an ICC for a measure with a low between-subject variance will be low, even if error variance is also low, so the use of this statistic in homogeneous populations is of limited value (Mehta et al., 2018).

Exploratory analyses allowed us to look instead at error variance alone, by examining absolute within-subject test-retest differences. These were low: less than 1.2 dB, on average, in both test languages and environments. This gives promise that Arabic thresholds measured online will be no less reliable than those measured in the lab and/or in other languages. To confirm this assertion, it is essential to collect data from a more diverse sample with considerable between-subject DIN threshold variance.

It's worth noting that retest DIN thresholds were slightly lower (by ~0.4 dB, on average) than test thresholds, across both languages and environments. This may indicate a small training effect, which is consistent with other studies that reported improvements of up to 1.5 dB at retest (Jansen et al., 2010; Smits et al., 2013; Vlaming et al., 2014; Koifman et al., 2016). The fact that the training effect was relatively small in our data may be due to specific features of the online listening software: the inclusion of clear instructions, demos, performance feedback, and an initial practice block.

Arabic and English DIN thresholds and audiometric thresholds

Similar to the ICC results above, the informative value of our correlations with audiometric thresholds is limited by the homogeneity of the study sample. All participants had clinically normal hearing, leading to low between-subject variance in both PTA and DIN thresholds, and low correlations between the two. Several studies have reported strong associations between DIN thresholds (in various languages) and standard audiometric thresholds, but have done so using cohorts of listeners varying widely in hearing loss severity (Folmer et al., 2017; Jansen et al., 2010; Koole et al., 2016; Potgieter et al., 2018; Watson et al., 2012). Jansen et al. (2010) reported results closer to our own: weak correlations between French DIN thresholds and audiometrically normal hearing thresholds (less than 25 dB HL). Clearly, the inclusion of participants with a range of audiometric profiles would be informative in understanding the relation between PTA and Arabic DIN thresholds.

EHF thresholds were similarly uncorrelated with both English and Arabic DIN thresholds. This outcome is unsurprising, not only due to sample homogeneity but also the filtering of the DIN stimuli below 8 kHz. Our results are consistent with those of Shehabi et al. (2023), who found no significant association between EHF thresholds and English DIN thresholds (with similar spectral characteristics) in audiometrically normal participants.

Limitations

A crucial limitation of the study is restriction of the sample to normal-hearing participants, which precluded meaningful comparison of audiometric and DIN thresholds and assessment of DIN reliability. Validation in individuals with hearing loss crucial for any DIN test material intended for hearing screening, since adults with hearing loss are the primary target population. Recruitment of a large sample of audiometrically diverse Arabic speakers should provide insights into the effectiveness of the Arabic DIN in distinguishing hearing-impaired and normal-hearing populations, and define appropriate cut-off values for this task. Greater between-subject variance in DIN thresholds should yield interpretable measures of reliability. A less obvious benefit is that validation in a more diverse population may reveal variability in performance due to cultural and linguistic factors that might not have been evident in our limited sample – an important additional test of real-world validity.

The other key limitation of the study is the use of non-optimized digits. It is well established that digits within the range 1-10 typically vary in intelligibility when presented at the same SNR. Failure to correct for this inherent variability leads to flatter intelligibility functions for the test as a whole, reducing the statistical precision of the DIN threshold measurement and/or extending test duration (Zokoll et al., 2012). A solution is to measure the psychometric functions of individual digit tokens and, for each, determine the SNR at which a given scoring target is achieved (e.g., 80% correct). The levels of the digits can then be adjusted accordingly, homogenizing the intelligibility of the digits at each nominal SNR (Akeroyd et al., 2015).

In the present study, digits were not optimized, likely adding noise to the adaptive tracks. This limitation is less pernicious in this preliminary evaluation study than it would be in a screening test, since the research questions primarily concern differences between measurement conditions, all of which employed the same non-optimized stimuli. However, it is reasonable to ask whether variability in the adaptive tracks was within acceptable limits. It is somewhat reassuring that the mean SD of the turnpoints (averaged across both lab-based and at-home sessions) was 2.08 dB, i.e., similar to the step size of the adaptive track. However, optimization of the digits prior to their use in hearing screening is a clear and pressing priority. Moreover, the DIN thresholds gathered in the present study should not serve as normal-hearing reference values when interpreting DIN data gathered using optimized digit stimuli in future; fresh validation data should be collected, from both hearing-impaired and normal-hearing listeners.

Conclusions

This study provides promise that the browser-based DIN is capable of producing remotely measured data that are reasonably comparable with gold-standard lab DIN data. It also suggests that using Arabic rather than English stimuli has effects that are statistically significant but modest. Nonetheless, it is wise

to consider systematic differences across test material and test environments when comparing freshly measured DIN thresholds with previously acquired reference data. The browser-based DIN has also been shown to produce small test-retest differences, regardless of language or environment. However, the use in this study of only audiometrically normal participants resulted in low between-subject variability in DIN, PTA, and EHF thresholds, leading to poor test-retest reliability of DIN thresholds and weak correlations between PTA and DIN data. Further research including more audiometrically diverse participants is warranted, as is optimization of the digits to homogenize their intelligibility.

Acknowledgments

The study was funded by the School of Health Sciences at the University of Manchester, the Medical Research Council (MR/V01272X/1), and the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre (BRC) (NIHR203308). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. We would like to thank all our participants for their time and commitment to the study.

Data availability statement

The datasets presented in this study can be found online in an Open Science Framework repository (https://osf.io/4wg58/files/osfstorage/65965bfa0344617c3868f563).

References

Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., Gagné, J. P., Lutman, M., Wouters, J., Wong, L., Kollmeier, B., & International Collegium of Rehabilitative Audiology Working Group on Multilingual Speech Tests (2015). International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests. ICRA Working Group on Multilingual Speech Tests. *International Journal of Audiology*, *54 Suppl 2*, 17–22. https://doi.org/10.3109/14992027.2015.1030513

British Society of Audiology (2013). Tympanometry. British Society of Audiology, Reading, UK. https://www.thebsa.org.uk/guidance-and-resources/current-guidance/

British Society of Audiology (2018). Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking. British Society of Audiology, Reading, UK. https://www.thebsa.org.uk/guidance-and-resources/current-guidance/

Buschermöhle, M., Wagener, K.C., Berg, D., Meis, M. & Kollmeier, B. (2014). The German digit triplets test (Part I): implementations for telephone, internet and mobile devices. Z. Für Audiol., 53, 139–145.

Buschermöhle, M., Wagener, K.C., Berg, D., Meis, M. & Kollmeier, B. (2015). The German digit triplets test (Part II): validation and pass/fail criteria. Z. Für Audiol., 54, 6–13.

Dillon, H., Beach, E. F., Seymour, J., Carter, L., & Golding, M. (2016). Development of Telscreen: a telephone-based speech-in-noise hearing screening test with a novel masking noise and scoring procedure. *International Journal of Audiology*, *55*(8), 463–471. https://doi.org/10.3109/14992027.2016.1172268

Folmer, R. L., Vachhani, J., McMillan, G. P., Watson, C., Kidd, G. R., & Feeney, M. P. (2017). Validation of a computer-administered version of the digits-in-noise test for hearing screening in the United States. *Journal of the American Academy of Audiology*, *28*(2), 161–169. https://doi.org/10.3766/jaaa.16038

Heinrich, A., Henshaw, H., & Ferguson, M. A. (2015). The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests. *Frontiers in Psychology*, *6*, 1–14. https://doi.org/10.3389/fpsyg.2015.00782

Hochmuth, S., Jürgens, T., Brand, T., Kollmeier, B. (2015) Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation? International Journal of Audiology, 54(Suppl. 2), 23-34

Hyvärinen, P., Fereczkowski, M., & MacDonald, E. N. (2023). Test-retest evaluation of a notched-noise test using consumer-grade mobile audio equipment. *International Journal of Audiology*. https://doi.org/10.1080/14992027.2022.2161955

Jansen, S., Luts, H., Wagener, K. C., Frachet, B., & Wouters, J. (2010). The French digit triplet test: A hearing screening tool for speech intelligibility in noise. *International Journal of Audiology*, *49*(5), 378–387. https://doi.org/10.3109/14992020903431272

Koifman, S., Buschermöhle, M., Holube, I., & Zokoll, M. A. (2016). Comparing evaluation data of the digit triplet test for Arabic, Hebrew, and Persian. *Deutschen Gesellschaft Für Audiologie 19. Jahrestagung*, 1–5. https://www.researchgate.net/publication/306129165

Koole, A., Nagtegaal, A. P., Homans, N. C., Hofman, A., De Jong, R. J. B., & Goedegebure, A. (2016). Using the digits-in-noise test to estimate age-related hearing loss. *Ear and Hearing*, *37*(5), 508–513. https://doi.org/10.1097/AUD.0000000000000282

Mehta, S., Bastero-Caballero, R. F., Sun, Y., Zhu, R., Murphy, D. K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. Statistics in Medicine, 37(18), 2734–2752. https://doi.org/10.1002/sim.7679

Ozimek, E., Kutzner, D., Sek, A., & Wicher, A. (2009). Development and evaluation of Polish digit triplet test for auditory screening. *Speech Communication*, *51*(4), 307–316. https://doi.org/10.1016/j.specom.2008.09.007

Potgieter, J. M., Swanepoel, D. W., Myburgh, H. C., Hopper, T. C., & Smits, C. (2016). Development and validation of a smartphone-based digits-in-noise hearing test in South African English. *International Journal of Audiology*, *55*(7), 405–411. https://doi.org/10.3109/14992027.2016.1172269

Potgieter, J. M., Swanepoel, D. W., & Smits, C. (2018). Evaluating a smartphone digits-in-noise test as part of the audiometric test battery. *South African Journal of Communication Disorders*, *65*(1), 1–6. https://doi.org/10.4102/sajcd.v65i1.574

Shehabi, A. M., Prendergast, G., Guest, H., & Plack, C. J. (2023). Binaural temporal coding and the middle ear muscle reflex in audiometrically normal young adults. *Hearing Research*, *427*, 1–12. https://doi.org/10.1016/j.heares.2022.108663

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. http://doi.org/10.1037/0033-2909.86.2.420

Smits, C. (2016). Comment on 'International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests', by Akeroyd et al. *International Journal of Audiology*, *55*(4), 268–269. https://doi.org/10.3109/14992027.2015.1131339

Smits, C., Goverts, T. S., & Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *The Journal of the Acoustical Society of America*, *133*(3), 1693–1706. https://doi.org/10.1121/1.4789933

Smits, C., & Houtgast, T. (2007). Recognition of digits in different types of noise by normal-hearing and hearing-impaired listeners. *International Journal of Audiology*, *46*(3), 134–144. https://doi.org/10.1080/14992020601102170

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology*, *43*, 15–28. https://doi.org/10.1080/14992020400050004

Smits, C., Merkus, P., & Houtgast, T. (2006). How we do it: The Dutch functional hearing-screening tests by telephone and internet. *Clinical Otolaryngology*, *31*(5), 436–440. https://doi.org/10.1111/j.1749-4486.2006.01195.x

Smits, C., Watson, C. S., Kidd, G. R., Moore, D. R., & Goverts, S. T. (2016). A comparison between the Dutch and American-English digits-in-noise (DIN) tests in normal-hearing listeners. *International Journal of Audiology*, *55*(6), 358–365. https://doi.org/10.3109/14992027.2015.1137362

Van den Borre, E., Denys, S., van Wieringen, A., & Wouters, J. (2021). The digit triplet test: a scoping review. *International Journal of Audiology*, *60*(12), 946–963. https://doi.org/10.1080/14992027.2021.1902579

Vlaming, M. S. M. G., Mackinnon, R. C., Jansen, M., & Moore, D. R. (2014). Automated Screening for High-Frequency Hearing Loss. *Ear & Hearing*, *35*, 667–679.

Watson, C. S., Kidd, G. R., Miller, J. D., Smits, C., & Humes, L. E. (2012). Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a US version. *Journal of the American Academy of Audiology*, *23*(10), 757–767. https://doi.org/10.3766/jaaa.23.10.2

Wolmarans, J., De Sousa, K. C., Frisby, C., Mahomed-Asmail, F., Smits, C., Moore, D. R., & Swanepoel, D. W. (2021). Speech recognition in noise using binaural diotic and antiphasic digits-in-noise in children: Maturation and self-test validity. *Journal of the American Academy of Audiology*, *32*(5), 315–323. https://doi.org/10.1055/s-0041-1727274

World Health Organization (2018). Addressing the rising prevalence of hearing loss. https://apps.who.int/iris/handle/10665/260336

World Health Organization (2013). Multi-Country Assessment of National Capacity to Provide Hearing Care. https://iris.who.int/handle/10665/339286

World Health Organization (2023). Deafness and Hearing Loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Zadeh, L. M., Silbert, N. H., Sternasty, K., & Moore, D. R. (2021). Development and validation of a digits-in-noise hearing test in Persian. *International Journal of Audiology*, *60*(3), 202–209. https://doi.org/10.1080/14992027.2020.1814969

Zokoll, M. A., Wagener, K. C., Brand, T., Buschermöhle, M., & Kollmeier, B. (2012). Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization prototype. *International Journal of Audiology*, *51*(9), 697–707. https://doi.org/10.3109/14992027.2012.690078