Laryngoscope
Investigative Otolaryngology

ORIGINAL RESEARCH OPEN ACCESS

Reliability Analysis of the Arabic Speech Matrix Test

Alia Asiri[1] | Fida Almuhawas[1,2] | Dalal Alrushaydan[3] | Mada Aljabr[3] | Tamer A. Mesallam[2] | Medhat Yousef[1,4]

[1]King Abdullah Ear Specialist Center (KAESC), King Saud University Medical City, Riyadh, Saudi Arabia | [2]Otolaryngology Department, College of Medicine, King Saud University, Riyadh, Saudi Arabia | [3]Cochlear Arabia Regional Headquarter, Riyadh, Saudi Arabia | [4]Audiovestibular Unit, ENT Department, Faculty of Medicine, Menoufia University, Menoufia, Egypt

Correspondence: Medhat Yousef (medhatf78@yahoo.com)

Received: 12 October 2024 | **Revised:** 24 January 2025 | **Accepted:** 25 March 2025

Keywords: hearing impairment | reliability testing | speech in noise test | speech matrix tests

ABSTRACT

Objective: Speech matrix tests offer information about a person's capacity to comprehend speech in noisy environments, which is an essential component of everyday communication, in contrast to pure tone audiometry, which primarily assesses hearing sensitivity. This study aimed to assess the test–retest reliability of the Arabic Speech Matrix test.

Methods: This is a prospective cohort study that included three groups: normal hearing individuals, cochlear implant users, and those using hearing aids. Seventy-five participants were included in the study. The test was administered in two different settings with noise presented from various angles. The test was re-administered to participants after a 7–14 days interval, and Intra-class Correlation Coefficient (ICC) and Bland–Altman plots were used to evaluate reliability.

Results: Moderate to excellent reliability was demonstrated, with higher consistency observed among hearing-impaired groups using cochlear implants and other devices. Minor learning effects were noted in the normal hearing group, with better reliability observed in the left setting.

Conclusion: The Arabic Speech Matrix test demonstrated strong test–retest reliability overall, indicating that it can be successfully incorporated into regular clinical audiological evaluations.

Level of Evidence: 4

1 | Introduction

Audiological assessments are essential for the evaluation and diagnosis of hearing impairment, with pure tone audiometry (PTA) as the principal tool for measuring hearing sensitivity. Nonetheless, while the PTA evaluates an individual's ability to detect pure tones, it doesn't comprehensively reflect the complexities of speech perception, especially in challenging environments [1]. Pure tone audiometry, although helpful for assessing fundamental hearing thresholds, fails to include the cognitive challenges necessary for evaluating speech recognition in noisy environments, which is crucial for daily communication [2, 3]. On the other hand, adaptive speech-in-noise assessments provide a more comprehensive and reliable approach for evaluating speech recognition challenges in realistic settings [4–8].

Individuals with hearing impairment, particularly those using hearing aids or implanted devices, often have challenges in comprehending speech, since speech intelligibility is more complex than simple signal detection [9]. To more accurately represent real-world communication, speech stimuli need to be included in audiological assessments [10]. Currently, several speech assessments rely on a clinician's voice, potentially introducing inconsistency in findings owing to discrepancies in tone, gender, and accent between the examiner and the patient [11].

Alia Asiri and Fida Almuhawas have contributed equally to this work.

The Arabic Speech Matrix test [12], a pre-recorded speech-in-noise assessment, overcomes those limitations by providing consistent criteria. It simulates daily situations using structured phrases created from a list of 50 words categorized into five groups: name, verb, number, noun, and adjective. The test is very flexible, producing up to 100,000 distinct phrases, rendering memorization by patients difficult. This facilitates repeated testing without compromising the results, and its short duration (about 4 min per list) improves clinical efficiency [13]. This test is crucial for evaluating patients' performance in auditory environments with background noise and for determining the efficacy of hearing aids in practical settings. The Arabic speech matrix may be administered to individuals aged 12 and older, since several linguistic skills are acquired by that age. This assessment is crucial for evaluating patients' performance in auditory environments with background noise and for determining the efficacy of hearing aids in practical settings [14, 15].

Reliability, defined as the consistency of a test in measuring a variable, can be assessed through several methods including internal consistency, parallel forms, interrater reliability, and test–retest reliability [16]. Despite the significant need for matrix sentence tests, no research has yet explored the reliability of the Arabic Speech Matrix. Therefore, this study aims to investigate its test–retest reliability for potential applications in clinical audiology. The Arabic matrix test will offer a valuable tool for obtaining reproducible, efficient, and internationally comparable speech recognition data for native Arabic-speaking individuals. It has significant potential for clinical use, research applications, and various hearing rehabilitation purposes.

2 | Methodology

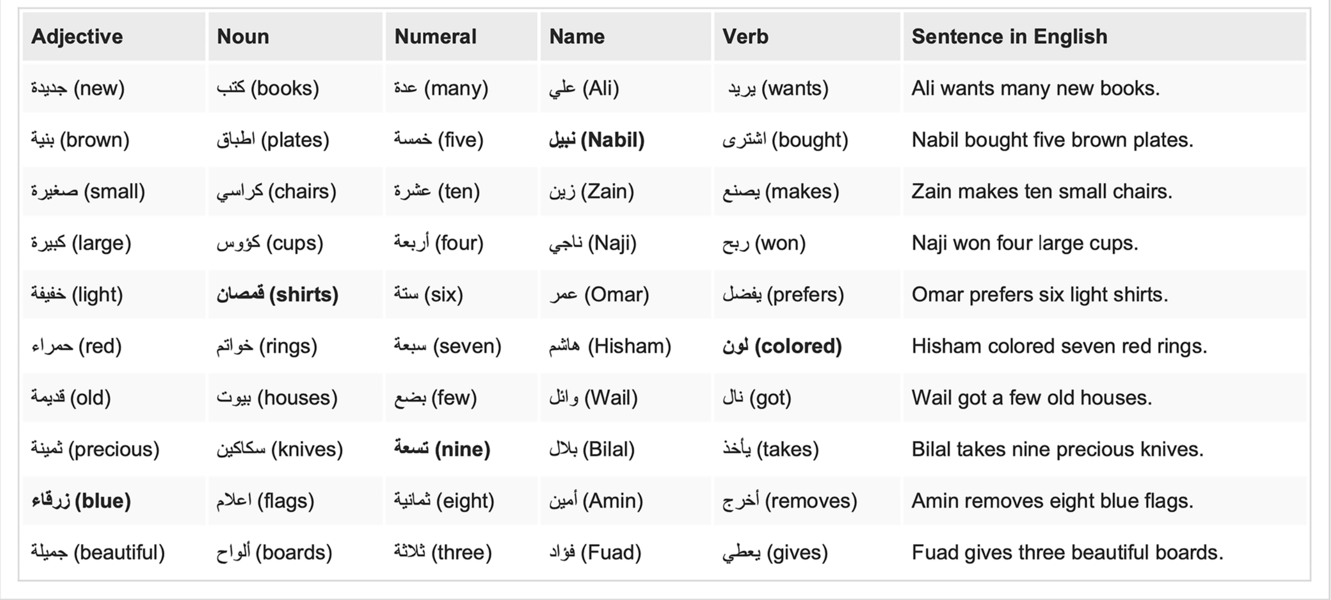
This prospective cohort study was approved by the Institutional Review Board at King Saud University's College of Medicine (IRB Number E-21-6437). A total of 75 participants were enrolled, including 31 normal hearing individuals, 31 bilateral cochlear implant recipients, and 13 bilateral hearing aid users. Inclusion criteria required participants to be over 12, have normal cognitive function, and be native Arabic speakers. Hearing device users had to have an aided word recognition score (WRS) of 65% or higher at a conversational level.

2.1 | Test Setting

Participants with normal hearing completed a comprehensive audiological evaluation, which included pure tone audiometry across frequencies from 250 Hz to 8000 Hz. Sound field aided assessment was carried out for hearing aid and cochlear implant users. A single 10-item training session preceded the Arabic speech matrix test, which was administered in two different settings. In setting one, speech was presented from the front (0°) with narrow band noise coming from the right (+90°), while in setting two, the noise came from the left (−90°). Participants were instructed to repeat each sentence after hearing it. The sentences were structured with five words (verb, name, numeral, noun, adjective) such as "يعطي فؤاد ثلاثة ألواح جميلة" (Figure 1). Each correctly repeated word was awarded one point, with a qualified audiologist marking the correct responses. The Arabic speech matrix test was repeated after approximately 2 weeks, using distinct test lists for each participant in each evaluation.

2.2 | Statistical Analysis

The primary reliability measure used was the Intra-class Correlation Coefficient (ICC), which assesses the ratio of variance between assessments to the total variance. ICC values below 0.5 indicate poor reliability, values between 0.5 and 0.75 suggest moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values above 0.9 represent excellent reliability, based on a 95% confidence interval. Additionally,



Adjective	Noun	Numeral	Name	Verb	Sentence in English
جديدة (new)	كتب (books)	عدة (many)	علي (Ali)	يريد (wants)	Ali wants many new books.
بنية (brown)	اطباق (plates)	خمسة (five)	**نبيل (Nabil)**	اشترى (bought)	Nabil bought five brown plates.
صغيرة (small)	كراسي (chairs)	عشرة (ten)	زين (Zain)	يصنع (makes)	Zain makes ten small chairs.
كبيرة (large)	كؤوس (cups)	أربعة (four)	ناجي (Naji)	ربح (won)	Naji won four large cups.
خفيفة (light)	**قمصان (shirts)**	ستة (six)	عمر (Omar)	يفضل (prefers)	Omar prefers six light shirts.
حمراء (red)	خواتم (rings)	سبعة (seven)	هاشم (Hisham)	**لون (colored)**	Hisham colored seven red rings.
قديمة (old)	بيوت (houses)	بضع (few)	وائل (Wail)	نال (got)	Wail got a few old houses.
ثمينة (precious)	سكاكين (knives)	**تسعة (nine)**	بلال (Bilal)	يأخذ (takes)	Bilal takes nine precious knives.
زرقاء (blue)	اعلام (flags)	ثمانية (eight)	أمين (Amin)	أخرج (removes)	Amin removes eight blue flags.
جميلة (beautiful)	ألواح (boards)	ثلاثة (three)	فؤاد (Fuad)	يعطي (gives)	Fuad gives three beautiful boards.

FIGURE 1 | Base matrix of the Arabic matrix sentence test [12]. A sample sentence is highlighted in bold, with each base sentence accompanied by its English translation in the rightmost column.

Pearson correlations were calculated for comparison purposes. While a strong correlation may suggest a relationship, it does not guarantee high reliability. To visually assess inter-assessment agreement, Bland–Altman plots were used, showing the difference between two measurements against their mean [17]. The limits of agreement, defined as ±2 standard deviations from the mean difference, were plotted to demonstrate the range within which 95% of differences would fall, assuming a normal distribution. Spaghetti plots were also used to compare assessments.

3 | Results

This study involved 75 participants, comprising 40 females and 35 males, aged between 18 and 70 years (mean age: 32±13.9 years). Table 1 outlines the demographic characteristics of each study group. Table 2 presents the Arabic speech matrix test results for the three groups: normal hearing individuals (group 1), cochlear implant recipients (group 2), and hearing aid users (group 3), along with the mean differences between the two assessments and their 95% confidence intervals. Higher reliability is indicated by narrower confidence intervals, and acceptable reliability requires a grand mean close to zero, with individual differences falling within ±2 standard deviations.

Table 3 presents the findings of the ICC. The ICC is an indicator of the reliability of ratings for clusters. Notably, ICC exhibited overall strong concordance across all groups. In the normal-hearing group, the right-ear setting showed moderate reliability with an ICC of 0.72, while the left-ear setting exhibited good reliability at 0.83. Groups of cochlear implant recipients and hearing aid users exhibited excellent reliability in both settings, reinforcing the robustness of the assessments across different auditory devices.

TABLE 1 | The demographic characteristics of each study group.

Group	Number of participants	Sex (F—female, M—male)	Age range	Mean of age and standard deviation
Normal hearing	31	20 F, 11 M	20–57	35±8
Cochlear implant recipients	31	15 F, 16 M	13–70	29.3±17.6
Hearing aid users	13	5 F, 8 M	13–68	31.2±15.4
Total	75	51 F, 45 M	13–70	32±13.9

TABLE 2 | Overview of right and left settings measurements for the three groups.

Group	Statistics	Right 1st	Right 2nd	Diff right 1–2	Left 1st	Left 2nd	Diff left 1–2
Normal	*n*	31	31	31	31	31	31
	Mean (SD)	−10.7 (2.98)	−12.3 (2.71)	1.6 (1.57)	−12.1 (2.75)	−12.8 (2.85)	0.6 (1.53)
	95% CI	−11.8, −9.6	−13.3, −11.3	1.0, 2.2	−13.1, −11.1	−13.8, −11.8	0.0, 1.1
	Median	−10.7	−12.4	1.6	−11.8	−12.5	0.1
	Range	−17 to −5	−18 to −7	−3 to 6	−17 to −7	−19 to −7	−3 to 4
Cochlear implant	*n*	31	31	31	31	31	31
	Mean (SD)	14.9 (15.39)	15.6 (19.73)	−0.7 (14.50)	13.9 (16.45)	12.5 (14.41)	1.4 (4.98)
	95% CI	9.0, 20.7	8.1, 23.1	−6.2, 4.8	7.6, 20.1	7.0, 18.0	−0.5, 3.3
	Median	11.9	12.0	1.2	10.7	16.6	0.6
	Range	−10 to 43	−10 to 80	−72 to 12	−11 to 39	−11 to 32	−6 to 18
Hearing aids	*n*	13	13	13	13	13	13
	Mean (SD)	7.2 (9.66)	4.9 (10.82)	1.1 (3.19)	4.7 (13.13)	5.2 (12.68)	−0.5 (5.70)
	95% CI	1.0, 13.3	−1.6, 11.5	−0.9, 3.1	−3.2, 12.7	−2.5, 12.9	−3.9, 3.0
	Median	4.7	2.8	0.8	3.5	3.0	0.7
	Range	−4 to 27	−9 to 26	−5 to 7	−10 to 38	−8 to 30	−13 to 8

Note: Right 1st: The first assessment with noise coming from right side. Right 2nd: The second assessment with noise coming from right side. Diff right 1–2: Difference between first and second assessment with noise coming from right side. Left 1st: The first assessment with noise coming from left side. Left 2nd: The second assessment with noise coming from left side. Diff left 1–2: Difference between first and second assessment with noise coming from left side. *n*: number. 95% CI: 95% confidence intervals.

Figure 2A illustrates the Bland–Altman plot for the normal-hearing group, right settings. It reveals two data points deviating from the ± 2 SD limits. The mean difference is notably below zero, with the mean line at −1.6, indicating moderate reliability. Figure 2C displays the spaghetti plot for the same group, where the second assessment consistently yielded lower values than the first assessment except for four participants. Similarly, Figure 2B shows the Bland–Altman plot for the left setting in the same group. The mean difference was at −0.56, indicating good reliability between the two assessments. Figure 2D presents the corresponding spaghetti plot.

Figure 3A–C demonstrate the Bland–Altman plot and the spaghetti plot for right setting in the cochlear implant group. All

TABLE 3 | ICC (Intra-Class coefficients) for the three groups.

Group 1: Normal hearing participants		
Measure	**ICC**	**Agreement**
Noise from right side	0.71614	Moderate
Noise from left side	0.83394	Good
Group 2: CI recipients		
Noise from right side	0.94336	Excellent
Noise from left side	0.94597	Excellent
Group 3: Hearing aid users		
Noise from right side	0.94816	Excellent
Noise from left side	0.90906	Excellent

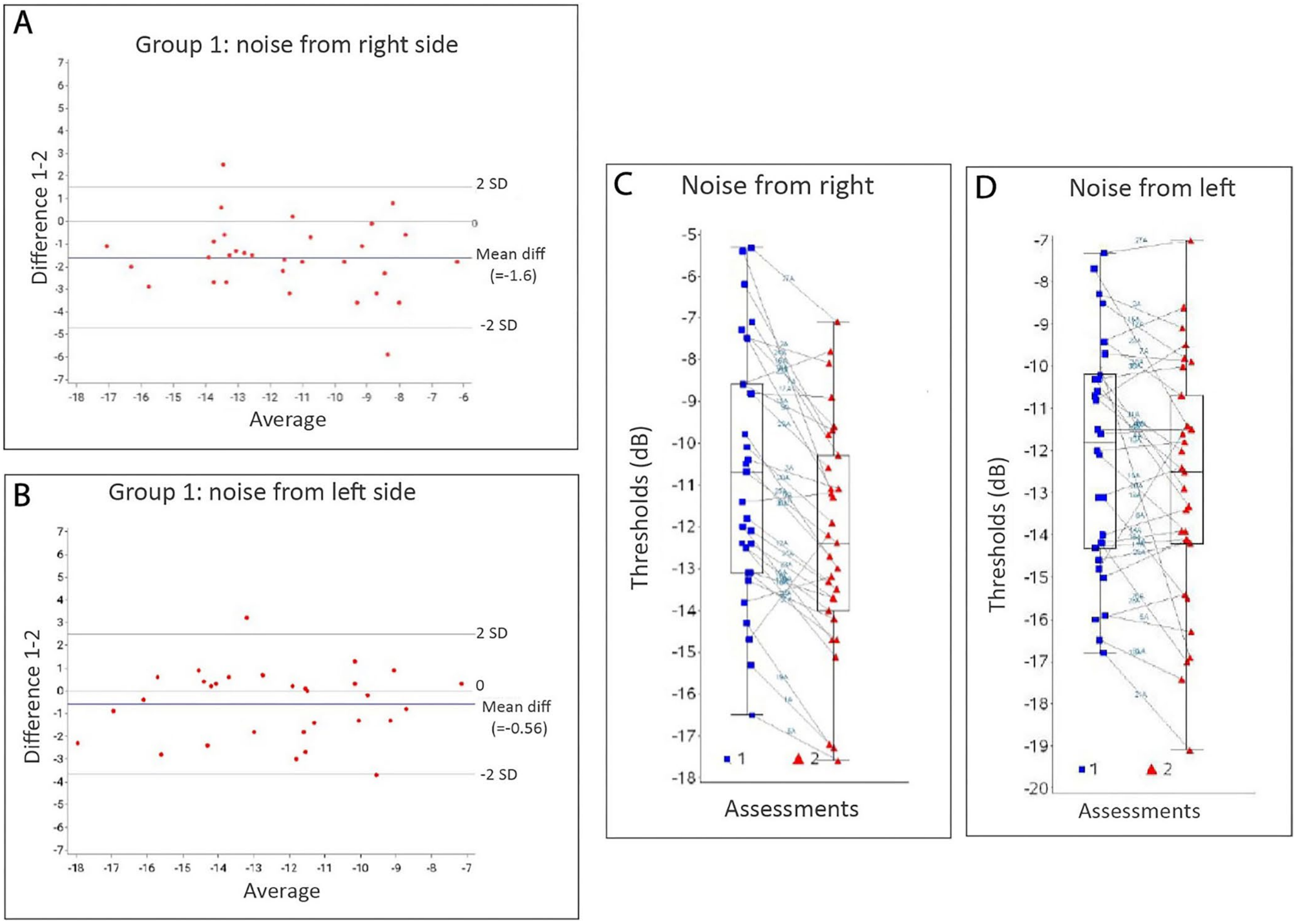


FIGURE 2 | (A) Bland–Altman plot for the first group with speech from front and noise from right side (mean diff: mean difference, SD: standered deviation). (B) Bland–Altman plot for the first group with speech from front and noise from left side. (C) Spaghetti plot for the first group with speech from front and noise from right side (dB: decibel). (D) Spaghetti plot for the first group with speech from front and noise from left side.

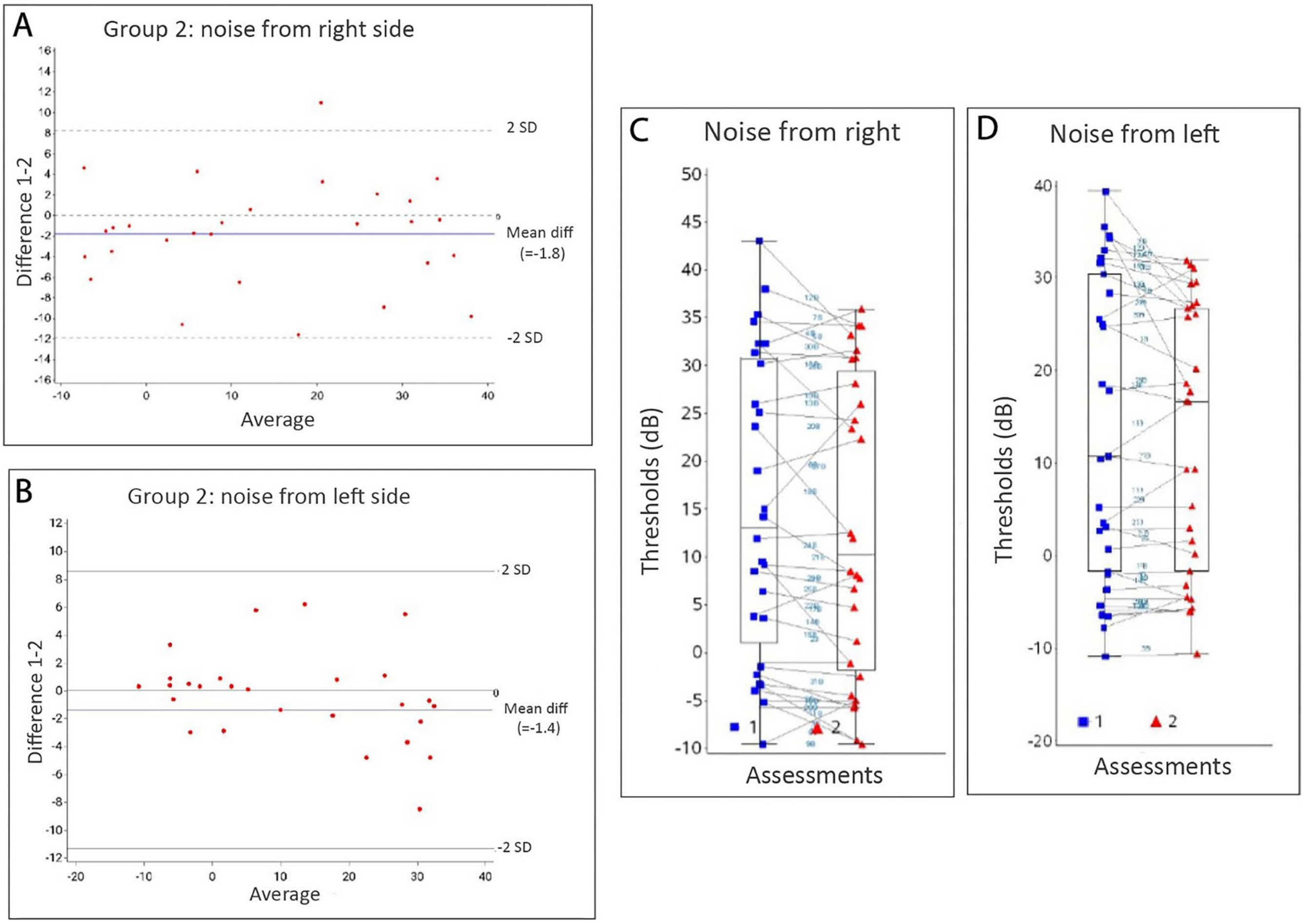


FIGURE 3 | (A) Bland–Altman plot for the second group with speech from front and noise from right side (mean diff: mean difference, SD: standered deviation). (B) Bland–Altman plot for the second group with speech from front and noise from left side. (C) Spaghetti plot for the second group with speech from front and noise from right side (dB: decibel). (D) Spaghetti plot for the second group with speech from front and noise from left side.

differences lie within the ±2 SD limits, indicating that the reliability is excellent. Figure 3B–D reflect the left setting of the same group. There are minor differences between the two assessments, a rather good correlation pointing to excellent reliability. All differences lie within the ±2 SD limits. The results of the hearing aid user group were represented in Figure 4 for both right and left settings. All points lie within the ±2 SD limits, which indicates overall excellent reliability.

4 | Discussion

The speech matrix test, recently included in hearing evaluations, has been translated into various languages, including Arabic. This study aimed to evaluate the test–retest reliability of the Arabic speech matrix test in both normal-hearing and hearing-impaired individuals [18]. The interval between test and re-test repetition must be long enough to avoid memory effects but short enough to prevent clinical changes [19]. Therefore, 75 participants were tested, with a retest interval of 7–14 days. The test was conducted across different hearing levels and devices, including cochlear implants and hearing aids.

Overall, all conducted tests demonstrated a satisfactory degree of reliability across all groups, with high consistency between both settings. Notably, in the normal hearing group, the left settings showed slightly better performance than the right, with a 0.12 difference in ICC values. The spaghetti plot for the left settings also revealed better thresholds in the second assessment, suggesting a possible learning effect, which was less pronounced in other groups. Previous studies, such as Puglisi et al. [20], similarly reported slight training effects when using the speech matrix test. Their findings, including a 1.5 dB training effect in initial measurements, closely align with our results for the first group. Additionally, Primadita's [21] study on the Indonesian matrix test found that two training lists were needed to ensure accurate results, even with high SNR.

The Arabic speech matrix test showed high reliability, correlating well with outcomes from various international matrix assessments, including the Finnish and European variants [22]. To ensure accurate measurements, it is essential to incorporate the Arabic matrix test into comprehensive audiological evaluations. Its results should be analyzed in conjunction with other audiological assessments. Furthermore, the implementation of two training lists and providing clear instructions to the patients before the main measurement are crucial steps to enhance the test's reliability and applicability in clinical settings.

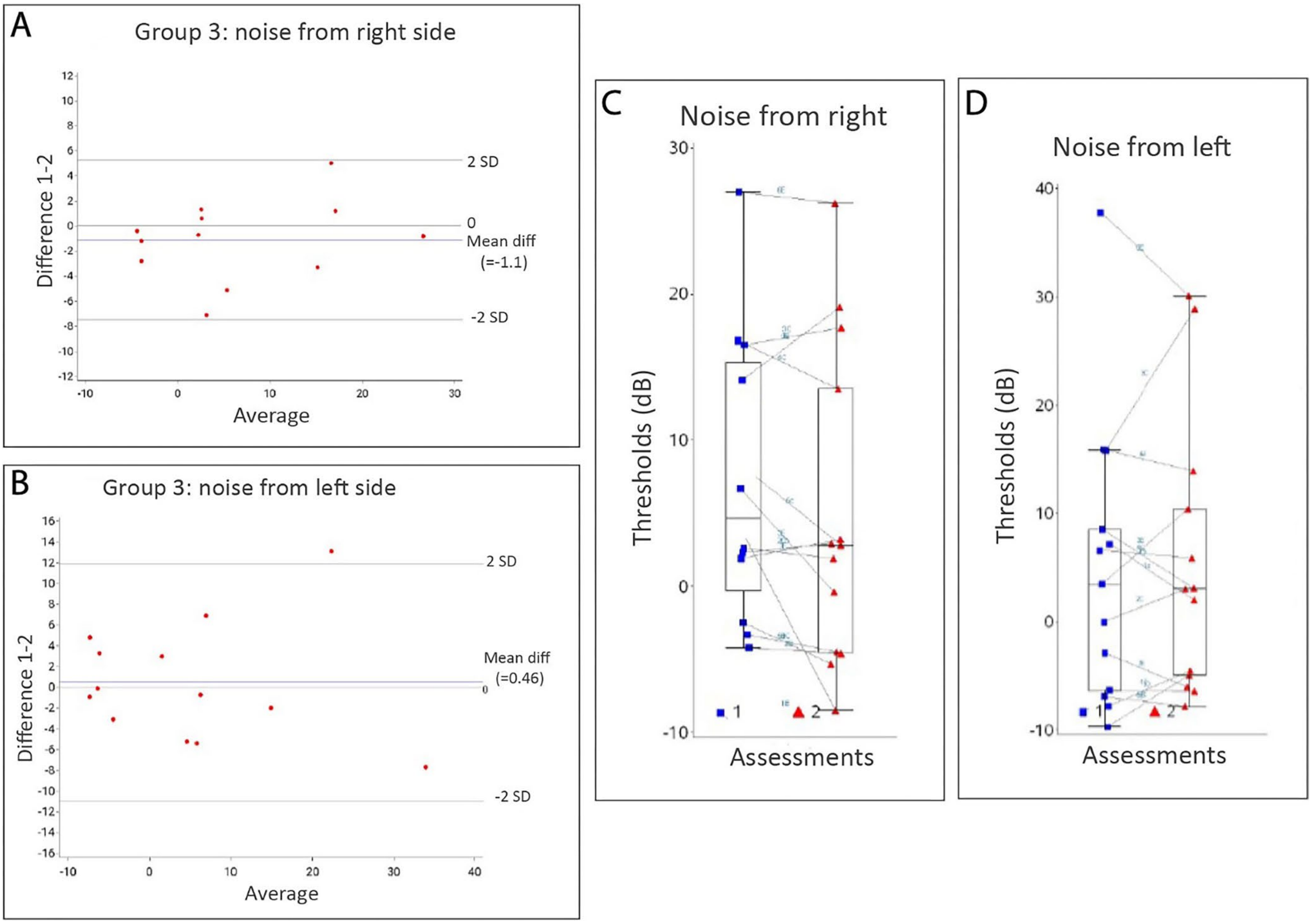


FIGURE 4 | (A) Bland–Altman plot for the third group with speech from front and noise from right side (mean diff: mean difference, SD: standered deviation). (B) Bland–Altman plot for the third group with speech from front and noise from left side (dB: decibel). (C) Spaghetti plot for the s third group with speech from front and noise from right side. (D) Spaghetti plot for the third group with speech from front and noise from left side.

5 | Conclusion

The Arabic speech matrix test showed strong test–retest reliability across all settings. To alleviate the learning effect, administering two training lists is essential. These findings validate the inclusion of the Arabic speech matrix test as a valuable tool in audiological evaluations, reinforcing its reliability and suitability for clinical use as part of the standard audiological test battery.

Acknowledgments

The authors are grateful to the Deanship of Scientific Research, king Saud University, for funding through the Vice Deanship of Scientific Research Chairs.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. M. Decambron, F. Leclercq, C. Renard, and C. Vincent, "Speech Audiometry in Noise: SNR Loss per Age-Group in Normal Hearing Subjects," *European Annals of Otorhinolaryngology, Head and Neck Diseases* 139, no. 2 (2022): 61–64.

2. C. Smits, C. S. Watson, G. R. Kidd, D. R. Moore, and S. T. Goverts, "A Comparison Between the Dutch and American-English Digits-In-Noise (DIN) Tests in Normal-Hearing Listeners," *International Journal of Audiology* 55, no. 6 (2016): 358–365.

3. M. A. Zokoll, K. C. Wagener, T. Brand, M. Buschermohle, and B. Kollmeier, "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, no. 9 (2012): 697–707.

4. K. C. De Sousa, W. Swanepoel, D. R. Moore, H. C. Myburgh, and C. Smits, "Improving Sensitivity of the Digits-In-Noise Test Using Antiphasic Stimuli," *Ear and Hearing* 41, no. 2 (2020): 442–450.

5. G. A. Miller, G. A. Heise, and W. Lichten, "The Intelligibility of Speech as a Function of the Context of the Test Materials," *Journal of Experimental Psychology* 41, no. 5 (1951): 329–335.

6. R. Plomp, "A Signal-To-Noise Ratio Model for the Speech-Reception Threshold of the Hearing Impaired," *Journal of Speech and Hearing Research* 29, no. 2 (1986): 146–154.

7. I. Ramkissoon, A. Proctor, C. R. Lansing, and R. C. Bilger, "Digit Speech Recognition Thresholds (SRT) for Non-Native Speakers of English," *American Journal of Audiology* 11, no. 1 (2002): 23–28.

8. C. Smits, S. E. Kramer, and T. Houtgast, "Speech Reception Thresholds in Noise and Self-Reported Hearing Disability in a General Adult Population," *Ear and Hearing* 27, no. 5 (2006): 538–549.

9. F. Martin and J. G. Clark, "Speech Audiometry," in *Introduction to Audiology*, 10th ed., ed. F. Martin and J. G. Clark (Allyn and Bacon, 2009).

10. F. Leclercq, C. Renard, and C. Vincent, "Speech Audiometry in Noise: Development of the French-Language VRB (Vocale Rapide Dans le Bruit) Test," *European Annals of Otorhinolaryngology, Head and Neck Diseases* 135, no. 5 (2018): 315–319.

11. Association AS-L-H, "Calibration of Speech Signals Delivered via Earphones," *ASHA* 29, no. 6 (1987): 44–48.

12. M. A. Zokoll, MB, N. Abdulhaq, S. Saleh, et al., "Applicability and Normative Data for an Arabic Matrix Sentence Test for Speech Recognition in Noise," *Cureus* 17, no. 1 (2025): e77062, https://doi.org/10.7759/cureus.77062.

13. HörTech, "InternatIonal MatrIx Tests," 2023.

14. R. E. Owens, *Language Development: An Introduction* (Allyn & Bacon, 1996).

15. R. E. Owens, *Language Development: An Introduction* (Pearsson, 2012).

16. F. Middleton, "The 4 Types of Reliability in Research | Definitions & Examples," 2022.

17. J. M. Bland and D. G. Altman, "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement," *Lancet* 1, no. 8476 (1986): 307–310.

18. A. C. Souza, N. M. C. Alexandre, and E. B. Guirardello, "Psychometric Properties in Instruments Evaluation of Reliability and Validity," *Epidemiol Serv Saude* 26, no. 3 (2017): 649–659.

19. C. B. Terwee, S. D. Bot, M. R. de Boer, et al., "Quality Criteria Were Proposed for Measurement Properties of Health Status Questionnaires," *Journal of Clinical Epidemiology* 60, no. 1 (2007): 34–42.

20. G. E. Puglisi, A. Warzybok, S. Hochmuth, et al., "An Italian Matrix Sentence Test for the Evaluation of Speech Intelligibility in Noise," *International Journal of Audiology* 54, no. Suppl 2 (2015): 44–50.

21. F. Primadita, *Development and Clinical Validation of the Indonesian Matrix Sentence Test: Faculty of Medicine and Health Sciences > Department of Medical Physics and Acoustics* (Carl von Ossietzky Universität Oldenburg, 2021).

22. A. Dietz, M. Buschermohle, A. A. Aarnisalo, et al., "The Development and Evaluation of the Finnish Matrix Sentence Test for Speech Intelligibility Assessment," *Acta Oto-Laryngologica* 134, no. 7 (2014): 728–737.