Supplementary Table 1: Detailed Speech Recognition Percentages, Signal-to-Noise Ratios (SNRs), and Performance Criteria

Study	Assessment measures	Speech in Quiet /noise	Measurement (percentage / SNR) & Performance criteria
Gatehouse, (1992)	WRS & FAAF	WRS: Q FAAF: N	WRS used AB (S) word lists, difficulty equated British PB lists with ten items each. Masking noise simulated the speech spectrum in the non-test ear. SIN employed an adaptive FAAF test with 80 items, presenting them at a calibrated intensity of 65 dB SPL. In both adaptive and fixed SNR presentations, performance was assessed by randomizing the order of 80 items and measuring the percent correct identification of words.
Silman et al. (1993)	SRT, W-22, SIN & NST	SRT, NST, W-22: Q SIN: N	SRT testing covered taped spondaic W-1 words and suprathreshold speech-recognition with CID W-22 monosyllabic PB words. The SPIN test, with 96 sentences and 12-talker babble, employed an up-down adaptive procedure to find the SNR for 50 percent sentence recognition, presented at 40 dB SL. The NST consisted of seven subtests with a closed-set format.
Bentler et al, (1993a)	NST & SPIN	NST: Q/N SPIN: N	The SPIN test used a background of 12-talker babble at a +8 dB signal-to-noise ratio. It comprises 50 sentences, with half having high predictability and half low predictability. The NST utilized 62 items organized into subtests of seven to nine syllables, covering major consonant sounds with the vowels. Each subject heard two randomly selected NST lists per session, one in quiet and one in speech-weighted noise. Two types of noise were used: multi-talker babble and random noise. Presentation levels and signal-to-background noise ratios were constant for all sessions (+8 for SPIN, +5 for NST), with slight variations based on circuit activation or listener preference.
Cox et al., (1996)	CST & SPAC	N	The CST has 10 sentences per passage on common topics, spoken by a female talker, with six-talker babble as competing sound. Each passage has 25 scoring words, and this study used 12 passages (300 words) per score. The SPAC test requires selecting a target word from a set of four embedded in short sentences, featuring eight speech features, 48 items per form, and 12 test items per feature pair. To control for talker intelligibility, the same talker recorded material for both tests. Scoring for both CST and SPAC is based on percent correct responses.
Horwitz et al., (1997)	NST	N	Speech recognition ability was objectively assessed using a digitized subset of UCLA's NST stimuli. Subjects completed three randomized runs, each with different lists of 16 syllables varying in talker and consonant position, with the first run serving as a practice session and excluded from scoring. Scores were derived from four repetitions of the three lists (192 syllables in total). Speech was presented at 70 dB SPL with a background of speech-shaped noise at 50 dB SPL, determined through pilot studies to minimize ceiling effects while maintaining a sufficient signal-to-noise ratio. For subjects whose unaided practice scores exceeded 80%, the noise level was increased to 55-65 dB SPL to avoid ceiling effects.
Suanders & Cienkowski. (1997)	SRT-Q, PSRT-N, and SSRT-N.	SRT-Q PSRT-N SSRT-N	SRT-Q used CID W-1 spondees with a 2 dB step size. For SRT-N, HINT test sentences were employed, consisting of 250 short, intelligible sentences divided into 25 lists. Sentences were masked by spectrally shaped noise. SRT-Ns were measured through two procedures: Subjective SRT-N (SSRT-N), which eliminated silent intervals, and Performance SRT-N (PSRT-N), involving sentence repetition. S/N adjustments were made based on responses, with different SRT calculations for each procedure. Both SSRT-N and PSRT-N maintained the same speech level.
Munro & Lutman, (2003)	FAAF for quiet, normal, and raised speech levels	N	The FAAF test consists of 20 sets of four words, creating an 80-item vocabulary. Speech and noise levels were defined at the reference point, approximating quiet, normal, and raised speech at 55-, 62-, and 69-dB SPL, respectively. Subjects were tested at a fixed SNR across conditions, determined individually for each subject using an adaptive FAAF strategy targeting 71% correct at a 62 dB SPL presentation level. Performance was measured at this fixed SNR for all speech presentation levels, assessing the percentage of correctly identified words
Amorim & Almeida., (2007)	Speech discrimination	Q	Speech discrimination percentage scores were evaluated using a word list featuring 50 monosyllables recorded on a compact disk. The list was presented at a comfortable level of 25-40 dBNS above the SRT. This phonetically balanced word list consists of 25 monosyllable words presented in four different combinations. Each ear was individually assessed, totaling 50 words for the overall evaluation.
Choi et al., (2011)	WIN	N	The subjects were presented with 25 monosyllables in the Korean language, spoken live under multi-talker babble noise. The noise intensity was consistently set at 70 dB HL, with a SNR of -3 dB. The correct reaction rate was denoted as the speech discrimination score (SDS) in noise (%).
Song et al., (2011)	WRS	Q	WRS was assessed using 50 phonemically balanced words, presented through insert earphones at 30 to 40 dB above the patient's speech recognition threshold. Randomized presentation mitigated learning effects. Statistical significance, determined by Raffin and Thornton criteria, identified clinically significant differences between pre-aided and post-aided word recognition scores within the 70% to 100% range, while scores outside this interval were considered different.
Dawes et al., (2014a)	FAAF	N	The FAAF test comprises 20 sets of four binary words, resulting in an 80-item test that focuses on auditory/phonetic distinctions. Responses are scored as correct or incorrect, with an overall percentage and error breakdown. Target phrases were presented at 65 dB SPL and 75 dB SPL against steady noise with the same spectrum. At baseline, the SNR for 50% correct was determined for each participant and listening condition. Adaptive testing began with an initial noise level of 55 dB SPL (SNR +10 dB), with a 4 dB step size, reduced to 2 dB steps for the last ten reversals. The final eight reversals provided the SNR for the 50% criterion performance for each condition. This criterion SNR was then used for testing with the full 80-item FAAF version at baseline and week 12. Note that SNR was set independently for each condition.
Dawes et al., (2014b)	FAAF	N	The FAAF test comprises 20 sets of four binary words, totaling 80 items, scored for overall percentage correctness. Targets were presented at 65- and 75-dB SPL against a steady noise with the same spectrum, both from the same loudspeaker at zero degrees azimuth. In each session, participants first completed a familiarization run of 40 words at easy levels. At the baseline session, the adaptive FAAF determined the SNR for 50% correctness for each participant and listening condition. Subsequently, participants were tested at this SNR using the 80-item FAAF version at both baseline and week 12.
Lavie et al., (2015)	Dichotic Tests, SIN, Monosyllabic Word Identification in Quiet	Dichotic & SIN: N Mono syllabic: Q	Participants were tested on SIN and dichotic listening tasks using Hebrew word lists. For the SIN test, 28 lists of 20 bisyllabic words were presented with multitalker babble noise from three loudspeakers positioned around the listener. The SNRs were adjusted by varying the noise levels while keeping word levels constant, and SNRs were pre-identified for each participant to achieve 30%–80% word recognition. Follow-up tests at 1 and 2 months, as well as at a later stage, measured identification scores using the pre-identified SNRs to reflect both threshold estimation and everyday listening environments. For the dichotic listening test, participants were presented with six lists of 25 pairs of monosyllabic words, with one word of each pair presented to each ear simultaneously. The test aimed to determine ear dominance, with scores based on the percentage of correctly repeated words. The scores from the pretest established the dominant ear, which was used for further analysis.

Dawes & Munro., (2017)	SIN	N	Sentence were presented at 65 dB SPL from a center speaker at 0° azimuth and ear level, had initial background noise at an SNR of +4 dB. Six random noise recordings, free of intelligible speech, were used from ±45° azimuth loudspeakers, adjusting based on correct content word identification.The 50% correct recognition threshold was adaptively estimated, beginning with a 4 dB step size for the initial 4 reversals and decreasing to 2 dB steps for the final 10 reversals. The average of the last eight reversals determined the criterion performance SNR.
Habicht et al., (2018)	Processing times & SRT	SRT: N	The Oldenburg corpus consists of two sentence structures: subject–verb–object (low linguistic complexity) and object–verb–subject (high linguistic complexity), determined by subtle grammatical cues in German. Participants repeated words from these sentences, first in a training session of 40 sentences, then estimating the SNR for 80%-correct speech reception (SRT80) with 40 additional sentences. Stationary speech-shaped noise at 65 dB SPL was used for all measurements, with speech stimuli amplified based on individual REIGs.
Karawani et al. (2018b)	QuickSIN	N	QuickSIN is a nonadaptive clinical test where sets of six sentences were presented binaurally at 70 dB HL in four-talker babble. The sentences started at +25 dB SNR and decrease by 5 dB for subsequent sentences, reaching 0 dB SNR. SNR scores were averaged across four lists to get a composite score, with lower SNR loss values indicating improved speech understanding in noise.
Megha & Maruthy, (2020)	SIN	N	The study used 20 consonant-vowel syllables for the SIN test. These syllables were presented amid speech-spectrum shaped noise at 0 dB SNR to explore how participants adapted to task difficulty. Participants identified syllables from a randomly ordered set, with each syllable repeated five times.
(Wright & Gagné, (2021)	HINT	N	SNR, indicating the level for 50% sentence accuracy. The SNR for each sentence list was adjusted based on participant performance within specified score ranges (0-24%, 25-44%, 45-55%, 56-75%, 76-100%), resulting in proportional increases or decreases for subsequent lists.

Note: CID=Central Institute for the Deaf, CST= Connected Speech Test, DDT= Dichotic Digits Test, dB HL= decibels hearing level, dB SPL= decibels sound pressure level, dB SL= decibels sensation level, FAAF= Four Alternative Auditory Feature test, HINT= Hearing in Noise Test, N= Noise, NST= Nonsense Syllable Test, PB= Phonetically balanced, PSRT-N= Performance SRT-N, Q= Quiet, REIG= Real ear insertion gain, SNR= Signal to noise ratio, SRT= Speech recognition threshold, SPAC= Speech Performance in Noise, SPIN= Speech Perception in Noise, SSRT-N= Subjective SRT-N, WIN= Words in noise , WRS= Word Recognition Score.

Supplementary Table 2: Quality Assessment and Level of Evidence ratings

Study	12/14-question NIH quality assessment score	NIH Quality rating (good, fair, poor)	Level of evidence (CEBM)
Gatehouse (1992)	7/12	Fair	3
Silman et al. (1993)	10/14	Fair	3
Bentler et al. (1993a)	7/14	Fair	3
Bentler et al. (1993b)	7/14	Fair	3
Cox et al. (1996)	6/14	Fair	3
Horwitz et al. (1997)	7/14	Fair	3
Saunders & Cienkowski (1997)	7/14	Fair	3
Munro & Lutman (2003)	10/12	Good	3
Vestergaard (2006)	7/14	Fair	3
Amorim & Almeida (2007)	7/12	Fair	3
Metselaar et al. (2009)	9/14	Fair	2
Choi et al. (2011)	8/14	Fair	3
Song et al. (2011)	8/12	Fair	4
Dawes et al. (2013)	8/14	Fair	3
Dawes et al. (2014a)	9/14	Fair	3
Dawes et al. (2014b)	9/14	Fair	3
Lavie et al. (2015)	10/14	Fair	3
Chang et al. (2016)	5/14	Poor	3
Dawes & Munro (2017)	9/14	Fair	3
Habicht et al. (2018)	10/14	Fair	3
Karawani et al. (2018a)	11/14	Good	2
Karawani et al. (2018b)	11/14	Good	2
Megha & Maruthy (2020)	7/14	Fair	3
Wright & Gagné (2021)	13/14	Good	2
Karawani et al. (2022)	8/14	Fair	3