Trends in Hearing
Volume 27: 1–11
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165231191382
journals.sagepub.com/home/tia
Sage

BPACE: A Bayesian, Patient-Centered Procedure for Matrix Speech Tests in Noise

Christoph Schmid[1], Wilhelm Wimmer[1,2] and Martin Kompis[1]

Abstract

Matrix sentence tests in noise can be challenging to the listener and time-consuming. A trade-off should be found between testing time, listener's comfort and the precision of the results. Here, a novel test procedure based on an updated maximum likelihood method was developed and implemented in a German matrix sentence test. It determines the parameters of the psychometric function (threshold, slope, and lapse-rate) without constantly challenging the listener at the intelligibility threshold. A so-called "credible interval" was used as a mid-run estimate of reliability and can be used as a termination criterion for the test. The procedure was evaluated and compared to a STAIRCASE procedure in a study with 20 cochlear implant patients and 20 normal hearing participants. The proposed procedure offers comparable accuracy and reliability to the reference method, but with a lower listening effort, as rated by the listeners (−1.8 points on a 10-point scale). Test duration can be reduced by 1.3 min on average when a credible interval of 2 dB is used as the termination criterion instead of testing 30 sentences. Particularly, normal hearing listeners and well performing, cochlear implant users can benefit from shorter test duration. Although the novel procedure was developed for a German test, it can easily be applied to tests in any other language.

Keywords

speech in noise, listening effort, signal-to-noise ratio, psychometric function, adaptive procedure

Received 28 November 2022; Revised received 12 July 2023; accepted 14 July 2023

Introduction

Speech intelligibility tests in noise are essential instruments in the assessment and follow-up of patients treated with hearing aids and cochlear implants (CIs) (British Society of Audiology, 2019; Nilsson et al., 2011). Measuring a person's speech reception threshold (SRT) often requires a tradeoff between accuracy, test duration, and listener's effort. Longer test runs may provide more accurate results, but are time-consuming and can be tiring for the participant. Moreover, the listening effort related to being tested near the individual listener's intelligibility threshold may provoke stress (Zekveld et al., 2011; Mackersie & Cones, 2011).

Matrix sentence tests can be used to determine speech intelligibility in noise and offer a balance between accuracy, test duration and listening effort (Kollmeier et al., 2015). Hagerman (1982) describes how to form syntactically correct sentences out of a matrix of predefined words. These are then used to find a subject's SRT (e.g., in noise). To this end, an adaptive procedure alters the signal-to-noise ratio (SNR) depending on the listener's responses. Typically, the SNR is adjusted to reach 50% correct responses to determine the SRT in noise (i.e., the SNR that yields 50% intelligibility). Testing at this intelligibility level is efficient due to the steepness of the psychometric function (PF) (Green, 1990). However, many listeners feel uncomfortable when trying to repeat sentences they can barely understand. Furthermore, the SNRs corresponding to the SRTs may be lower than those encountered in everyday situations (Smeds et al., 2015; Wu et al., 2018) and may fall below the range of the SNRs where hearing aids' nonlinear features perform optimally (Naylor, 2016).

The German matrix sentence test Oldenburger Satztest (OLSA) applies 5-word-long sentences and requires 20–30 sentences to evaluate the SRT. To determine the SRT, it

[1] Department of Otorhinolaryngology, Head and Neck Surgery, Bern University Hospital, Inselspital, Bern, Switzerland
[2] Department of Otorhinolaryngology, TUM School of Medicine, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany

Corresponding Author:
Wilhelm Wimmer, Technical University of Munich, Germany; Department of Otorhinolaryngology, TUM School of Medicine, Klinikum rechts der Isar, Munich, Germany.
Email: wilhelm.wimmer@tum.de

alters the SNR in a so-called STAIRCASE procedure. In this procedure, the SNR is adapted in predefined steps according to the listener's responses (Wagener et al., 1999). While the procedure is simple and widely used, one of its limitations is that it is not suited for determination of the slope of the underlying PF. Knowing not only the SRT but also the slope would facilitate the identification of points of interest on the PF (e.g., SRT_{75}) and allow to estimate the effect of SNR changes on intelligibility (MacPherson & Akeroyd, 2014).

Besides the STAIRCASE procedure, there are other adaptive methods to determine the SRT in noise. Bayesian methods, for instance, combine prior knowledge and experimental results by using Bayes' theorem. This can be used to model a PF and infer its parameters based on prior values and listener responses. Herbert et al. (2022) simulated seven different adaptive procedures for speech-in-noise tests regarding the accuracy and reliability for SRT_{50} and SRT_{75}. They report promising results for Bayesian procedures and suggest to investigate these procedures with human participants.

The aim of our study was to design a procedure for matrix sentence tests in noise which requires less listening effort and thus offers more comfort to the participants. The method does not constantly challenge the listeners at their SRT, but tests at higher intelligibility levels too and uses as few trials as possible. To this end, Bayesian inference in an updated maximum likelihood (UML) procedure was used (Shen & Richards, 2012). This adaptive procedure has proved reliable in various psychophysical tests; for instance, in the field of psychoacoustics (Carcagno & Plack, 2021; Fischer et al., 2021a,b; Lee & Müllensiefen, 2020; Jurado et al., 2020). The method can estimate the reliability of the results by calculating a credible interval, the Bayesian equivalent to the confidence interval in frequentist statistics (Comotto, 2022). The credible interval can serve as the termination criterion for the adaptive procedure. The proposed method was called BPACE (Bayesian PAtient-CEntered) procedure since it focuses not only on efficiency but also on the patient's listening effort. Additionally, besides the determination of the SRT, the procedure can estimate the slope and the ceiling performance level of a listener's PF.

Methods

Test Procedure

Idea and Principle. The proposed test procedure assumes the speech intelligibility in noise to be a PF modeled by a logistic function of the SNR according to Equation 1, where α is a threshold parameter indicating the horizontal position of the function, β describes the function's slope, and λ is the distance of the upper asymptote to 100% correct. An example of a logistic function is plotted in Figure 1:

$$p = \frac{1-\lambda}{1+e^{-4\beta(SNR-\alpha)}}. \tag{1}$$

The lower bound of this function is zero, as used in the original version of the test Wagener et al. (1999) and as confirmed by a small pilot study. The parameter α corresponds to the SNR at the function's inflection point; for $\lambda = 0$ it is equal to the SRT. The PF converges to 1-λ for high SNRs. In the literature, the parameter λ is referred to as the lapse-rate and is often regarded as a nuisance parameter (Prins, 2013). For NH listeners, this holds true as their PF reaches one except for lapses in their attention. For listeners who do not reach 100% speech understanding in quiet, however, the term lapse-rate can be misleading since the upper bound of their PF is not limited by lapses only but by their ability to understand speech even under optimal conditions. For the sake of consistency, λ is nevertheless called lapse-rate in this publication.

Deducing a PF from previous responses allows testing subsequent sentences at SNR levels deviating from the currently assumed SRT. A UML procedure identifies optimal SNRs—so called sweet-points—for the next trial, based on the results of all previous trials. Testing the next sentence at one of those sweet-points will minimize the expected variance for the respective parameter and thus result in maximum information gain. Figure 1 shows an example of responses to trials at different SNRs and the estimated psychometric function. In this function, the sweet-points are illustrated. The α sweet-point is at the inflection point of the PF. Two points are needed to gain information about the slope, marked as lower and upper β sweet-points. For parameter λ there is strictly speaking not a single point that would minimize its variance, theoretically an infinitely high SNR would be ideal. In practise a high SNR can be taken as a proxy for the λ sweet-point.

After each trial, all the previous responses are evaluated to find the most probable PF.[1] Then the algorithm decides, which sweet-point on this PF to use for the next trial. Whenever less than 50% of the words were correctly repeated in the last trial, the algorithm selects a sweet-point with a higher SNR and vice versa. With the aim of not discouraging the listener, the lower β sweet-point was not used, even though this hampers the estimation of β. The sweet-point selection can thus be regarded as a 1-up, 1-down tracking rule between the α, the upper β and the λ sweet-point. This represents a compromise between SRT determination, listening effort and information gain on the PF's slope and asymptote.

The test can be terminated after a predefined number of trials or as soon as the reliability of the results is sufficient. At the end of the adaptive procedure, the first sentence is presented once more to the listener without noise to offer an easy finish, since Kahneman et al. (1993) found the end of a test contributing substantially to the participant's subjective experience. This sentence is neither evaluated nor saved.

Implementation. MATLAB R2020a (the Mathworks, Natick, MA, USA) was used to implement the test procedure

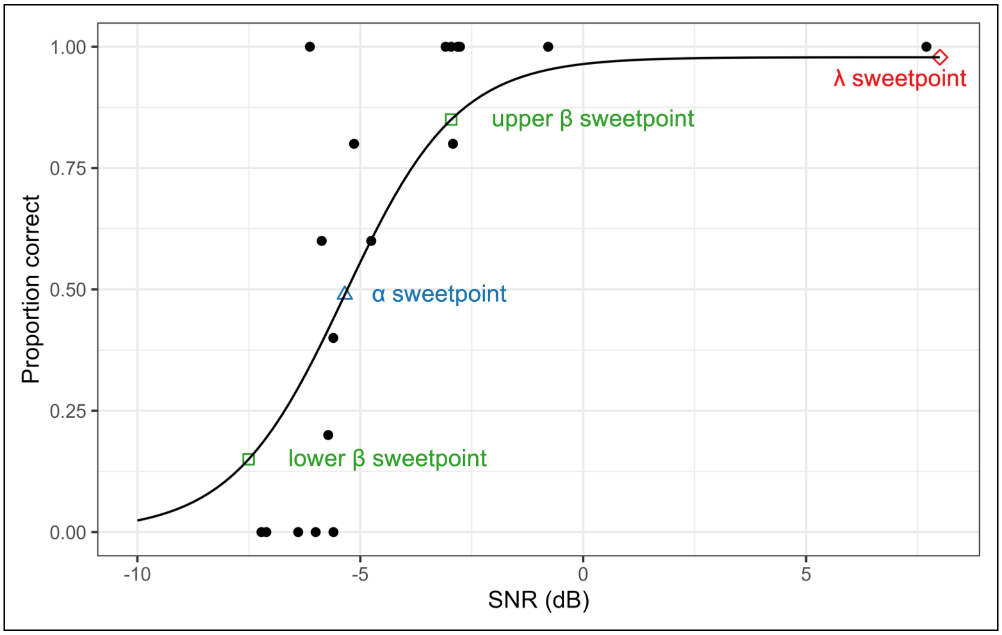

Figure 1. Listener's responses (black dots) depending on the signal-to-noise ratio (SNR) with the resulting psychometric function and its sweet-points (α as triangle, β as squares and λ as diamond).

algorithm with a graphical user interface. It stores the individual test steps as well as the test results in a database. The source code is available as Supplemental Material. It may be adapted to work with any matrix sentence test.

The UML toolbox version 4 provided by Shen et al. (2015) calculates the posterior parameter distribution, the sweet-points and the credible intervals. The procedure starts with normal prior distributions for each parameter according to Table 1. The prior distributions were defined and parametrized based on pilot tests and simulations with virtual listeners. For each parameter, a range with a number of uniformly distributed discrete values was defined (Table 1).

Together they span a three-dimensional parameter space in which the maximum likelihood for the posterior parameters is calculated using Bayes' theorem according to Equation 2 with ϕ denoting the parameter vector $\{\alpha, \beta, \lambda\}$ (Shen & Richards, 2012). $P(\phi_i|D)$ is the posterior probability for hypothesis ϕ_i given the data D and the prior probability $P(\phi_i)$. $P(D|\phi)$ is the probability for data D, given the parameter vector ϕ:

$$P(\phi_i|D) = \frac{P(D|\phi_i)P(\phi_i)}{\sum_j P(D|\phi_j)P(\phi_j)}. \quad (2)$$

Once the most likely parameter values are determined, the sweet-points on the new PF can be identified by minimizing the expected variance for each parameter. Equation 3 denotes the variance in general. Equations A1 to A3 by Shen & Richards (2012) show the specific variances for

Table 1. Parameter Space and Prior Distributions.

Parameter	α, $\mathrm{dB_{SNR}}$	β, $\mathrm{dB^{-1}}$	λ, –
Limits	−15; 20	0.05; 0.22	0; 0.4
Number of discrete values	140	40	40
μ	0	0.13	0
σ^2	10	0.07	0.15

α, β, and λ:

$$\sigma_\alpha^2 = p(SNR, \phi)[1 - p(SNR, \phi)]/\left(\frac{dp(SNR, \phi)}{d\alpha}\right)^2. \quad (3)$$

After every trial, the posterior distribution of each parameter is used to calculate the parameter's Bayesian credible interval. The interval shrinks narrower with an increasing number of trials with reliable responses. The 95% credible interval for the parameter α can serve as a reliability indicator or stop criterion, and the desired limit for test termination is adjustable in the software. We propose a limit of 2 dB as stop criterion, as the collected data show reliable results once this limit is reached. This limit is more strict than the ±1.1 dB standard deviation (*SD*) published for the OLSA by HörTech (2013) since for normally distributed data only 68% of the values lie within ±1 SD (i.e., 2.2 dB) in contrast to the desired 95% within the credible interval of 2 dB.

Technical Validation. We performed simulations to parametrize and validate our new procedure and to compare it to the conventional existing procedure. A detailed summary of the simulation experiments is provided in the Supplemental Material. Although simulations are useful to estimate the performance and accuracy of the different test methods, a study with real participants is required for a test validation.

Study Design and Ethics

The BPACE procedure was evaluated and compared to the original STAIRCASE procedure in a prospective study to test its accuracy and reliability. Since the novel procedure is parametrized with prior values, it is important to test different groups of listeners and therefore CI patients and normal hearing (NH) participants were included. The study was approved by the local institutional review board (BASEC-ID 2021-01828).

Participants. A total of 20 CI users and 20 NH listeners were evaluated in the study. All subjects were native German speakers and gave their informed consent. Pure tone thresholds for NH participants had to be 25 dB_{HL} or better. For CI users, speech intelligibility with their CI was required to be at least 65% when tested with monosyllabic words in quiet.

Demographics. The mean age of the study participants was 39 years (*SD* = 17 years); the CI patients had a mean of 47 years (*SD* = 21 years) and NH participants 32 years (*SD* = 6 years). A total of 27 participants were women and 13 men. The CI patients are implanted with Advanced Bionics, Cochlear and Med-El implants and have been using their systems for 1–24 years (mean 9 years, *SD* = 7 years). Details of the CI systems and the NH participant's pure tone thresholds are added as Supplemental Material.

Audiometric Setting. All experiments were performed in an acoustic chamber with a calibrated clinical audiometer (Equinox, Interacoustics A/S, Assens, Denmark). Pure tone air conduction hearing thresholds (in dB_{HL}) were measured for the NH participants at 500, 1.000, 2.000, 4.000, 6.000, and 8.000 Hz using headphones. For sound field measurements, a single loudspeaker for signal and noise was mounted at a distance of 1 m in front of the listener. For each sentence, a segment of the OLSA noise was tuned in with a raised-cosine 500 ms before the sentences started. This stationary, speech-shaped noise was generated by Wagener et al. (1999) through random superposition of the speech material. The maximum noise level was 85 dB_{HL}.

The CI patients used only the specified implant; in case of residual hearing in the contralateral ear, it was closed with an earplug.

Test Protocol. Figure 2 illustrates the study timeline, starting with an assessment of the prerequisites and a training list consisting of 20 sentences in quiet. Subsequently, two adaptive tests identified two SRTs for the first procedure (staircase or BPACE) before the listener was asked to rate the listening effort. The effort rating scale ranged from 0 to 10, with labels ranging between no effort to very high effort (Zekveld et al., 2011). Afterward, the same proceeding was repeated with the other test procedure. The order of the two test procedures (staircase/BPACE, A/B, respectively) was systematically alternated and not communicated to the listeners (single-blinded design).

To minimize learning effects, the participants were allowed to see 10 example sentences containing all possible words during the training session. However, the test setting was not closed-set, as the listeners could—and frequently did—give nonconforming responses or no response, thus the effective guessing rate was close to zero. For both test procedures, each test run included 30 sentences; even after seven trials at the same SNR with the STAIRCASE procedure, the test was not terminated.[2] The STAIRCASE method was adapting the signal level starting at 0 dB_{SNR} with a constant noise level of 65 dB_{HL}. For the BPACE method, on the other hand, the signal level was fixed at 65 dB_{HL} and the noise level adapted according to the responses. Thereby, the first sentence of the BPACE test could always be presented without noise, to facilitate the beginning for the listener at the expense of information gain. According to Wagener & Brand (2005), the resulting SRTs do not differ significantly when keeping the signal level constant instead of the noise level.

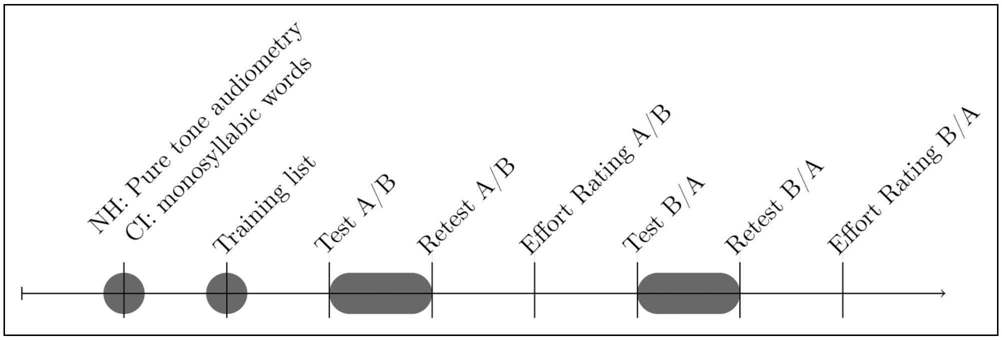

Figure 2. Timeline of the study.

Statistical Analysis

Demographic data and functional outcomes are summarized using descriptive statistics. For the BPACE method, the credible interval for parameter α is calculated after each sentence. As soon as it falls below the limit of 2 dB, the threshold is accepted as SRT. If the credible interval remains above the limit, the SRT after 30 sentences is evaluated. The STAIRCASE method is always evaluated after 30 sentences.

The agreement between the two methods is graphically explored in a Bland–Altman plot (Martin Bland & Altman, 1986). Furthermore, a linear mixed-effects model was used to investigate the differences between the SRT estimates obtained with the two test methods. The test method (i.e., STAIRCASE vs. BPACE), the test sequence number (i.e., 0–3), the hearing condition (i.e., NH vs. CI) and the age (years) were used as fixed effects. The subject ID was included as random intercept to account for repeated measurements. The subjective listening effort was evaluated with a similar model excluding the test sequence number. The residuals were visually inspected in a residuals versus fitted plot.

The test reliability was determined by calculating the intraclass correlation coefficient (ICC) using a two-way effects, absolute agreement, single measurement model. ICC can take values between 0 and 1 and is a measure of the consistency of test and retest results (< 0.5 poor, 0.5–0.75 moderate, 0.75–0.9 good, and > 0.9 excellent) (Koo & Li, 2016; Weir, 2005). The R studio software (Core Team, 2017) with the lme4 and irr packages (Bates et al., 2015; Gamer et al., 2019) served as tools for the statistical analysis.

Results

Accuracy and Agreement of the Two Methods

Figure 3 summarizes the SRTs measured with each test method for CI patients and NH participants separately. CI patients in this study showed a mean SRT of -1.3 dB$_{\mathrm{SNR}}$ ($SD = 2.1$ dB$_{\mathrm{SNR}}$) and the NH participants -6.6 dB$_{\mathrm{SNR}}$ ($SD = 0.8$ dB$_{\mathrm{SNR}}$). The NH participants hold a mean SRT of -6.6 dB$_{\mathrm{SNR}}$ ($SD = 0.7$ dB$_{\mathrm{SNR}}$) with the STAIRCASE method and -6.6 dB$_{\mathrm{SNR}}$ ($SD = 0.8$ dB$_{\mathrm{SNR}}$) with the BPACE method. For the CI patients, the mean SRT was -1.3 dB$_{\mathrm{SNR}}$ ($SD = 2.2$ dB$_{\mathrm{SNR}}$) with the STAIRCASEmethod and -1.4 dB$_{\mathrm{SNR}}$ ($SD = 2.0$ dB$_{\mathrm{SNR}}$) with the BPACE method.

Figure 4 demonstrates the learning curve over the four tests for NH and CI patients.

Only the BPACE method estimates the slope and the asymptote of the PF. For NH listeners, the mean slope was 0.18 dB^{-1} ($SD = 0.02$ dB^{-1}) and for CI patients 0.14 dB^{-1} ($SD = 0.03$ dB^{-1}). More details about the slope estimates are depicted in the supplemental section (Supplemental Figure S7).

The left panel in Figure 5 shows the Bland–Altman plot representing each test method by the mean of the two SRTs measured, subtracting STAIRCASE from BPACE results. In the right panel, the mean SRTs obtained with the BPACE method are plotted against the mean SRTs of the STAIRCASE method.

A linear mixed-effects model including the test method, test sequence number, the hearing condition, and age as fixed effects and the subject as random effect was computed to analyze the 160 data points. No significant dependence of the SRT on the test method (staircase or BPACE) was found

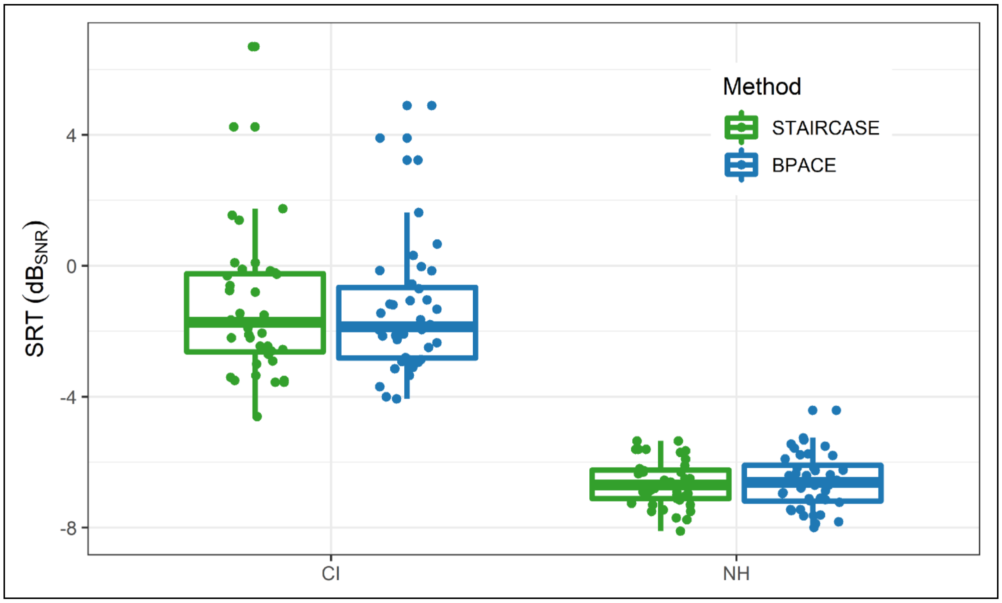

Figure 3. SRT for the STAIRCASE and the BPACE methods separated for CI patients and NH listeners. Abbreviations: SRT= speech reception threshold; CI = cochlear implant; NH = normal hearing; BPACE = Bayesian PAtient-CEntered.

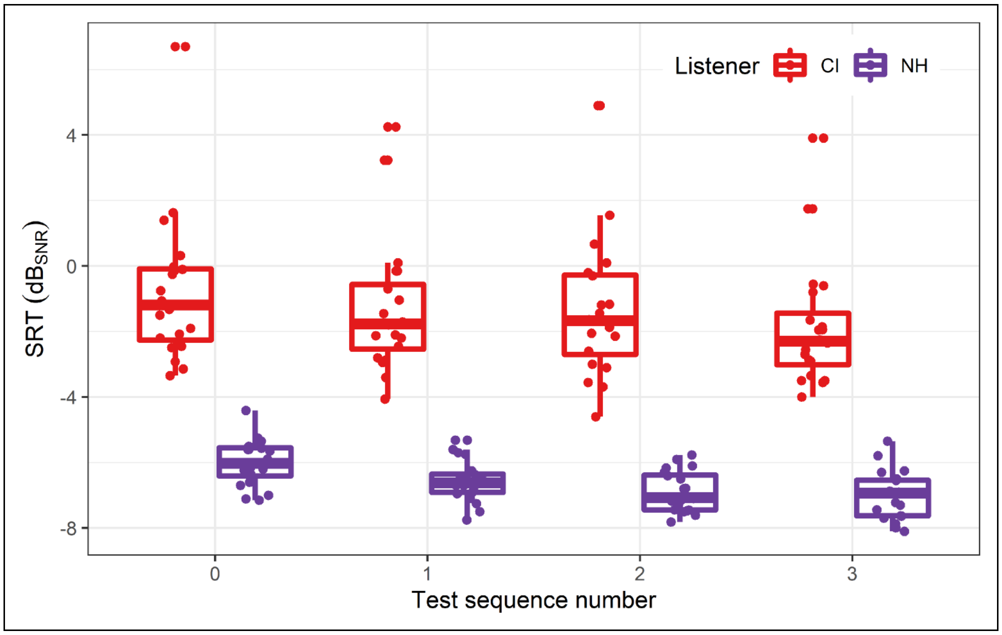

Figure 4. SRT depending on test sequence number for CI patients and NH listeners. Abbreviations: SRTs = speech reception thresholds; CI = cochlear implant; NH = normal hearing.

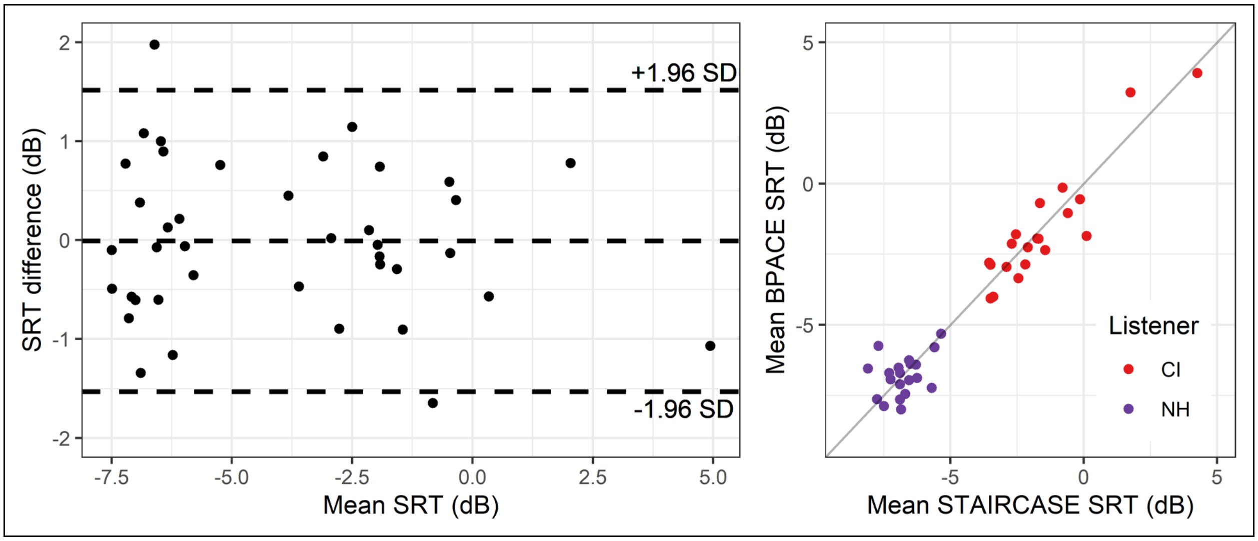

Figure 5. Left: Bland–Altman plot for STAIRCASE and BPACE test methods. Right: Scatterplot of the SRT, the color separates NH listeners from CI patients. Abbreviations: SRTs = speech reception thresholds; CI = cochlear implant; NH = normal hearing.

(see Table 2). Unsurprisingly, the NH group performed considerably better than the CI patients. In the course of the investigation, the SRT of the participants improved with each test run by 0.33 $\mathrm{dB_{SNR}}$ ($p < 0.001$). Age yielded a slightly positive effect, implying better results for younger participants. No specific pattern or outlier was detected in the residuals versus fitted plot.

Test–retest reliability was determined by calculating the ICC of the two measurements for each test method. The results are listed in Table 3. The absolute test–retest difference of the SRT was found to be below 1.1 dB for 33 out of 40 STAIRCASE test pairs and for 33 out of 40 BPACE test pairs.

Listening Effort

The subjective listening effort ratings and the proportion correct are summarized in Figure 6. The participants rated the effort with 6.4 (*SD* = 2) on the 10-point scale. For the STAIRCASE method, the rating was 7.3 (*SD* = 1.7) and for the BPACE method, it was 5.5 (*SD* = 1.9). The BPACE method reduced the mean subjective listening effort from 7.3 to 5.4 for the CI patients and from 7.4 to 5.6 for the NH participants. The proportion correct for the CI patients was 51.5% (*SD* = 2.8%) with the STAIRCASE method and 60.3% (*SD* = 3.9%) with the BPACE method. The NH listeners achieved a proportion correct of 53.8% (*SD* = 2.3%) with the STAIRCASE method and 63.2% (*SD* = 2.8%) with the BPACE method.

Table 4 gives a summary of the fitting of the linear mixed-effects model with method, age, and hearing condition as fixed effects and the subject as random effect. No specific pattern or outlier was detected in the residuals versus fitted plot.

Table 2. Linear Mixed-Effects Model Summary.

	Coefficient	95% confidence interval	*p*-value
Intercept	−3.04	[−4.44, − 1.63]	<0.001
Test method (BPACE)	0.01	[−0.16, 0.15]	0.926
Test sequence number	−0.33	[−0.40, − 0.26]	<0.001
Hearing condition (NH)	−4.61	[−5.52, − 3.69]	<0.001
Age (years)	.05	[0.02, 0.07]	0.001

Abbreviations: BPACE = Bayesian PAtient-CEntered; NH = normal hearing.

Test Duration

The mean test duration for 30 sentences in noise was 4.6 min (*SD* = 0.8 min)—4.8 min (*SD* = 0.8 min) with the STAIRCASE method and 4.4 min (*SD* = 0.7 min) with the BPACE method. Applying a termination criterion of 2 dB for the credible interval reduces the testing time to 3.1 min (*SD* = 1.2 min) for the BPACE method. The termination criterion was reached after 21.5 trials (*SD* = 6.3 trials) on average. Figure 7 compares the duration for 30 trials or for termination based on the running credible interval. The data are grouped for NH and CI listeners since the reduction in test duration is more pronounced for NH participants. For the training list in quiet, the mean duration was 2.5 min (*SD* = 0.6 min).

Discussion

Speech in noise testing is an important audiological outcome measure, however, testing time and listening effort limit its applicability in clinical routine. The BPACE procedure was designed to alleviate these drawbacks and still return reliable SRT estimates. The study results demonstrate very good agreement between the SRT estimates found by the STAIRCASE and BPACE methods. Figure 3 demonstrates almost equal

Table 3. Test–Retest Reliability.

	Normal hearing		Cochlear implant		All participants	
	STAIRCASE	BPACE	STAIRCASE	BPACE	STAIRCASE	BPACE
ICC	0.698	0.787	0.894	0.887	0.966	0.967

Abbreviations: BPACE = Bayesian PAtient-CEntered; ICC = intraclass correlation coefficient.

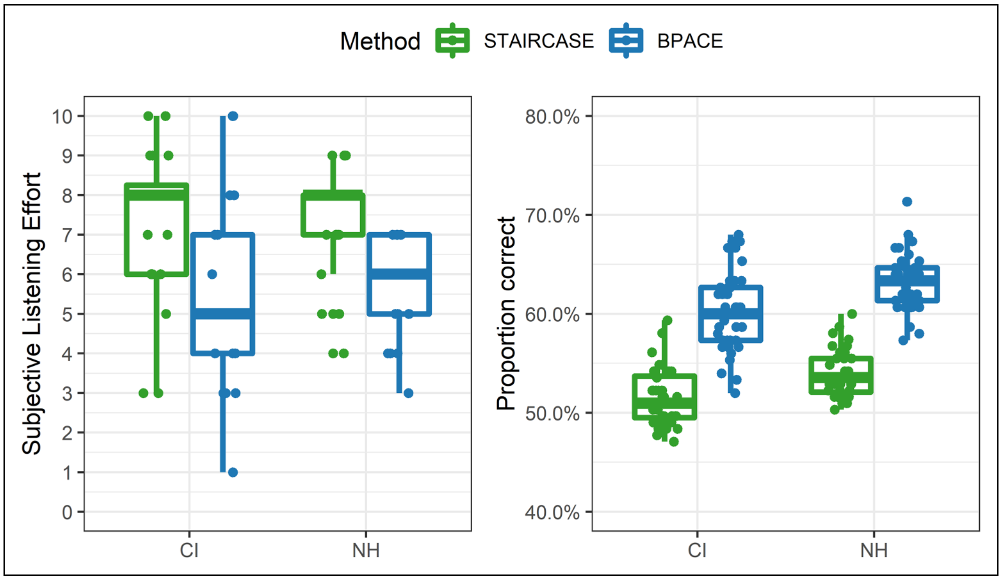

Figure 6. Subjective listening effort and mean proportion correct in the adaptive test for the CI patients and the NH participants. Abbreviations: CI = cochlear implant; NH = normal hearing.

mean SRTs for the two methods and comparable standard deviations. Figure 4 shows a decreasing SRT during the test sequence, a learning effect known from the literature (Wagener et al., 1999; Heyn, 2019; Rudolf & Kaiser, 2019). A longer training period with several training runs could account for this learning effect. However, in this study the learning effect is allocated equally to the STAIRCASE and to the BPACE results.

The mean SRT of $-6.6\,\mathrm{dB}_{\mathrm{SNR}}$ for the NH participants in this study lies between scores reported in the literature of $-7.1\,\mathrm{dB}_{\mathrm{SNR}}$ by Wagener et al. (1999) and $-6.3\,\mathrm{dB}_{\mathrm{SNR}}$ by Brand et al. (2004).

Besides providing the SRT in agreement with conventional existing procedures, the BPACE method yields more information about the entire PF. Knowing the parameters α, β, and λ permits the calculation of SRTs at more ecological SNRs (e.g., SRT_{75}) much more accurately than with the STAIRCASE method (Herbert et al., 2022). The BPACE method could also be adopted to test at a specific intelligibility level of particular interest instead of the sweet-points used.

Table 4. Linear Mixed-Effects Model Summary for the Listening Effort.

	Coefficient	95% confidence interval	*p*-value
Intercept	5.26	[3.67, 6.84]	<0.001
Test method (BPACE)	−1.82	[−2.34, − 1.31]	<0.001
Hearing condition (NH)	0.75	[−0.27, 1.77]	0.149
Age (years)	0.04	[0.01, 0.07]	0.006

Abbreviations: BPACE = Bayesian PAtient-CEntered; NH = normal hearing.

Accuracy and Agreement of the Two Methods

The mean SRT results for STAIRCASE and BPACE tests are almost equal. For CI patients, the mean and the standard deviation strongly depend on the selected patient collective, and it, therefore, cannot be compared to published data. For the NH participants, the SRT standard deviation was slightly higher for BPACE than for STAIRCASE tests and lower than the value reported for the OLSA, but in good agreement with other matrix sentence tests (HörTech, 2013; Kollmeier et al., 2015).

The Bland–Altman plot in Figure 5 and the linear mixed-effects model show a high agreement between the two test methods, suggesting that the SRTs obtained by the new method are valid. Moreover, the test–retest reliability is similar for both test methods.

The absolute test–retest difference of the SRT was found to be below 1.1 dB for 33 out of 40 tests for both methods. The credible interval calculated by the BPACE method would be able to detect five out of the seven results with more than 1.1 dB difference by virtue of their credible interval lying above the limit of 2 dB. With the STAIRCASE method, on the other hand, there is no possibility to discern the reliable results from those with a high test–retest difference.

Listening Effort

The CI patients and NH participants rated the listening effort for the BPACE method significantly lower than for the

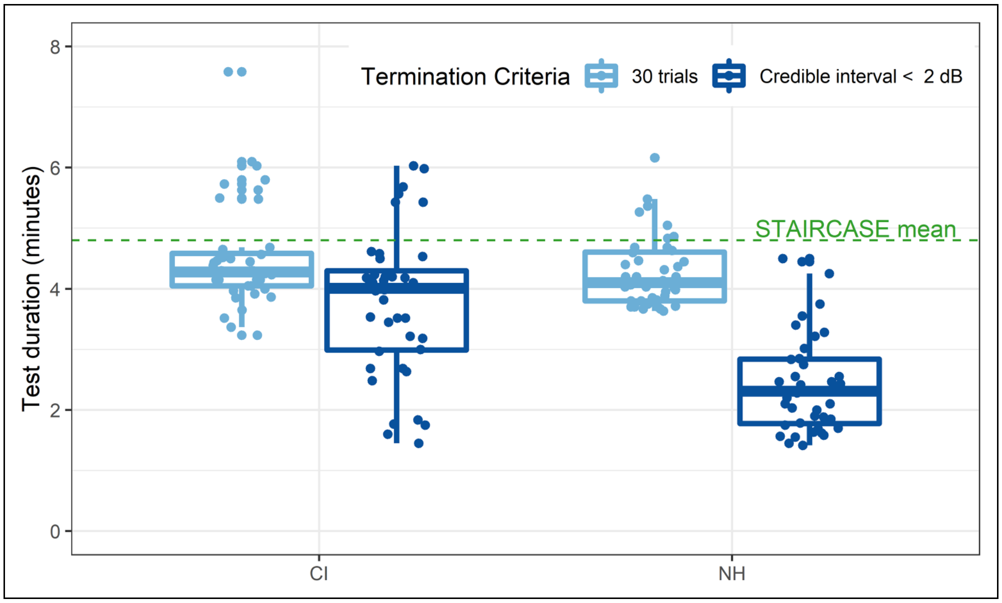

Figure 7. BPACE test duration for 30 trials or until reaching the desired credible interval for the CI patients and NH participants. The dashed line indicates the mean duration of the STAIRCASE method (4.8 min). Abbreviations: BPACE = Bayesian PAtient-CEntered; CI = cochlear implant; NH = normal hearing.

STAIRCASE method. It is worth mentioning that the better rating for the subjective listening effort is achieved by only a moderate increase in the overall proportion correct (see right graph in Figure 6). Apparently, for most listeners, it is already sufficient to get relief after a tough sentence and to not struggle at their limit permanently. Nonetheless, finding an SRT in noise remains a challenging task, and most participants rated the listening effort rather high. One NH participant criticized the noise level being too loud for the BPACE method and rated the listening effort higher than for the STAIRCASE method. This phenomenon could be avoided by keeping the overall level of signal and noise constant and changing the SNR by simultaneously raising one level while lowering the other (Kaandorp et al., 2014).

We did not find a significant influence of the SRTs on the effort ratings. Intuitively, one could suppose better performers to rate the listening effort lower. In particular, because the BPACE method effectively is more challenging for listeners with lower speech intelligibility in quiet. The upper asymptote of their PF is relatively low and, therefore, the α and β sweet-points are located at lower proportion correct levels. Thus, the overall proportion correct is lower for CI patients compared to NH listeners, as visible in the right graph of Figure 6. Interestingly, the STAIRCASE method generates the same effect even though this method converges at 50% proportion correct independently of the listener's performance. Most likely, the difference is caused by the first trials which are typically correct for NH listeners and incorrect for many CI patients while approaching the SRT starting at 0 dB SNR.

The linear mixed-effects model revealed a small but significant increase of the listening effort with the listener's age. This effect has been found and documented in previous studies (Cardin, 2016; Tun et al., 2009). However, the current work was not designed to investigate the influence of the listener's age on subjective listening effort.

Test Duration

For the BPACE method, a test termination based on the credible interval can substantially shorten the test for NH listeners. Unfortunately, only few CI patients reach an acceptable credible interval with few sentences and may thus benefit from this termination criterion. Poor performers with a shallow PF often require test runs with 30 sentences to produce reliable results. Therefore, Figure 7 demonstrates a smaller reduction of the test duration for CI patients compared to NH participants. Nevertheless, the testing time can be optimized individually when tests are terminated based on the updated credible interval.

The test duration with the STAIRCASE method was longer than with the BPACE method for 30 sentences. We assume the difference is caused by the overall difference in intelligibility and the listeners responded more rapidly to the more intelligible sentences presented at the upper β and the λ sweet-points.

It might be possible to define a termination criterion for the STAIRCASE procedure also, based on the variability or the reversals of the adapting SNR. Thereby, the test duration might be optimized, but such an advancement for the OLSA procedure is not in this project's scope. Termination after seven consecutive trials at the same SNR is another way to reduce the test duration but, according to experience, only rarely occurs. In this study, only two of the 80 STAIRCASE tests produced seven consecutive trials at equal SNR.

Limitations and Outlook

It is important to keep in mind that the calculated credible interval for the parameter α serves as an indicator for the SRT's reliability. But it is not the SRT's true credible interval for this depends on parameters β and λ too as well as the corresponding credible intervals.

Even though the slope and the asymptote of the PF can be estimated with the BPACE method, the accuracy of these values is limited. This is due to the sweet-point selection rule favoring the determination of the SRT. In contrast, there is no mathematical weighting of the parameters, as implemented in a different method proposed by Prins (2013).

Last, but not least, our evaluation is limited to NH participants and CI patients. It might be useful to evaluate the procedure with hearing aid users in a future study.

Conclusion

We developed a Bayesian procedure for matrix speech tests in noise. The method aims for reliable results in a short time with an acceptable listening effort and was compared to the existing procedure in a study. The evaluation shows good agreement of the novel test method BPACE with the conventional STAIRCASE method with similar test-retest reliability. The listeners rated the listening effort with the novel method significantly lower than for the STAIRCASE method. The estimation of a credible interval for the current SRT allows individual test termination. Thereby, the test duration can be reduced by 1.3 min on average, particularly for good performing listeners with consistent responses. The credible interval specifies the quality of the results, information not exploited by traditional methods.

Acknowledgments

We thank the two anonymous reviewers for their critical reading of the manuscript and the constructive comments. Thank you to Benjamin von Gunten and David Tschanz for their support in data collection.

Declaration of Conflicting Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was funded by the scientific fund of the ENT clinic Inselspital Bern. Open access funding was provided by the University of Bern.

ORCID iDs

Christoph Schmid https://orcid.org/0000-0003-0914-6244
Wilhelm Wimmer https://orcid.org/0000-0001-5392-2074

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The maximum likelihood method aims for the PF with parameters that maximize the likelihood of observing the given responses. It is described in more detail in the "Implementation" section.
2. The STAIRCASE procedure does not change the SNR if the listener responds two or three out of the five words correctly (after trial number five). For seven consecutive trials at the same SNR, most clinical applications terminate the test with the corresponding SNR as resulting SRT.

References

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brand, T., Wittkop, T., Wagener, K., & Kollmeier, B. (2004). Vergleich von Oldenburger Satztest und Freiburger Wörtertest als geschlossene Versionen. *DGA Jahrestagung*, (7),2–4. http://www.uzh.ch/orl/dga2004/programm/wissprogramm/Brand__T.pdf

British Society of Audiology. (2019). Practice Guidance Assessment of speech understanding in noise in adults with hearing difficulties. http://www.thebsa.org.uk/wp-content/uploads/2019/03/BSA-Practice-Guidance-Speech-in-Noise-FINAL.Feb-2019.pdf

Carcagno, S., & Plack, C. J. (2021). Effects of age on psychophysical measures of auditory temporal processing and speech reception at low and high levels. *Hearing Research*, *400*, 108117. https://doi.org/10.1016/j.heares.2020.108117

Cardin, V. (2016). Effects of aging and adult-onset hearing loss on cortical auditory regions. *Frontiers in Neuroscience*, *10*, 199. https://doi.org/10.3389/fnins.2016.00199

Comotto, F. (2022). Statistics 101: Credible vs Confidence Interval *Towards Data Science* Statistics 101 https://towardsdatascience.com/statistics-101-credible-vs-confidence-interval-af7b7e8fdd79

Core Team R. (2017). A language and environment for statistical computing.

Fischer, T., Schmid, C., Kompis, M., Mantokoudis, G., Caversaccio, M., & Wimmer, W. (2021). Effects of temporal fine structure preservation on spatial hearing in bilateral cochlear implant users. *The Journal of the Acoustical Society of America*, *150*(2), 673–686. https://doi.org/10.1121/10.0005732

Fischer, T., Schmid, C., Kompis, M., Mantokoudis, G., Caversaccio, M., & Wimmer, W. (2021). Pinna-imitating microphone directionality improves sound localization and discrimination in bilateral cochlear implant users. *Ear and Hearing*, *42*(1), 222–241.

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement.

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *The Journal of the Acoustical Society of America*, *87*(6), 2662–2674. https://doi.org/10.1121/1.399058

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, *11*(2), 79–87. https://doi.org/10.3109/01050398209076203

Herbert, N., Keller, M., Derleth, P., Kühnel, V., & Strelcyk, O. (2022). Optimised adaptive procedures and analysis methods for conducting speech-in-noise tests. *International Journal of Audiology*, 1–11. https://doi.org/10.1080/14992027.2022.2087112

Heyn, J. (2019). *Kurz- und langfristige Lern- und Habituationseffekte bei der Anwendung des Oldenburger Satztests an jungen, normalhörenden Probanden*. PhD Thesis, Universität Würzburg.

HörTech. (2013). *Bedienungsanleitung Oldenburger Satztest*. Oldenburg: Kompetenzzentrum für Hörgeräte-Systemtechnik.

Jurado, C., Larrea, M., Patel, H., & Marquardt, T. (2020). Dependency of threshold and loudness on sound duration at low and infrasonic frequencies. *The Journal of the Acoustical Society of America*, *148*(2), 1030–1038. https://doi.org/10.1121/10.0001760

Kaandorp, M. W., Smits, C., Merkus, P., Goverts, S. T., & Festen, J. M. (2014, January). Assessing speech recognition abilities with digits in noise in cochlear implant and hearing aid users. *International Journal of Audiology,* 54(1), 48–57. https://doi.org/10.3109/14992027.2014.945623

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54*(sup2), 3–16. https://doi.org/10.3109/14992027.2015.1020971

Koo, T., & Li, M. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lee, H., & Müllensiefen, D. (2020). The timbre perception test (TPT): A new interactive musical assessment tool to measure timbre perception ability. *Attention, Perception, & Psychophysics*, *82*(7), 3658–3675. https://doi.org/10.3758/s13414-020-02058-3

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113–122. https://doi.org/10.3766/jaaa.22.2.6

MacPherson, A., & Akeroyd, M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in Hearing*, *18*.2331216514537722. https://doi.org/10.1177/2331216514537722

Martin Bland, J., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Naylor, G. (2016). Theoretical issues of validity in the measurement of aided speech reception threshold in noise for comparing nonlinear hearing aid systems. *Journal of the American Academy of Audiology*, *27*(7), 504–514. https://doi.org/10.3766/jaaa.15093

Nilsson, M., McCaw, V., & Soli, S. (2011). Minimum Speech Test Battery for Adult Cochlear Implant Users. http://www.auditorypotential.com/MSTBfiles/MSTBManual2011-06-20.pdf.

Prins, N (2013). The PSI-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision*, *13*(7), 3. https://doi.org/10.1167/13.7.3

Rudolf, H., & Kaiser, M. (2019). *Habituation und Lerneffekte beim Oldenburger Satztest bei normalhörenden Probanden in Abhängigkeit vom Lebensalter*. PhD Thesis, Universität Würzburg.

Shen, Y., Dai, W., & Richards, V. M. (2015). A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure. *Behavior Research Methods*, *47*(1), 13–26. https://doi.org/10.3758/s13428-014-0450-6

Shen, Y., & Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *The Journal of the Acoustical Society of America*, *132*(2), 957–967. https://doi.org/10.1121/1.4733540

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology*, *26*, 183–196. https://doi.org/10.3766/jaaa.26.2.7

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*, *24*(3), 761–766. https://doi.org/10.1037/a0014802

Wagener, K., & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *International Journal of Audiology*, *44*(3), 144–156. https://doi.org/10.1080/14992020500057517

Wagener, K., Brand, T., Kuehnel, V., & Kollmeier, B. (1999). Entwicklung und evaluation eines Satztests in deutscher Sprache I–III: Design, Optimierung und Evaluation des Oldenburger Satztests. *Zeitschrift Für Audiologie*, *38*(3), 4–15.

Weir, J. P. (2005, Feb). Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res.*, *19*(1), 231–40. https://doi.org/10.1519/15184.1

Wu, Y., Stangl, E., Chipara, O., Hasan, S. A. W., & Oleson, J. (2018). Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild-to-moderate hearing loss. *Ear Hear*, *39*(2), 293–304. https://doi.org/10.1097/AUD.0000000000000486

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510.