Efficient Adaptive Speech Reception Threshold Measurements Using Stochastic Approximation Algorithms

Gertjan Dingemanse ⑩ and André Goedegebure

Abstract

This study examines whether speech-in-noise tests that use adaptive procedures to assess a speech reception threshold in noise (*SRT50n*) can be optimized using stochastic approximation (SA) methods, especially in cochlear-implant (CI) users. A simulation model was developed that simulates intelligibility scores for words from sentences in noise for both CI users and normal-hearing (NH) listeners. The model was used in Monte Carlo simulations. Four different SA algorithms were optimized for use in both groups and compared with clinically used adaptive procedures. The simulation model proved to be valid, as its results agreed very well with existing experimental data. The four optimized SA algorithms all provided an efficient estimation of the *SRT50n*. They were equally accurate and produced smaller standard deviations (SDs) than the clinical procedures. In CI users, *SRT50n* estimates had a small bias and larger SDs than in NH listeners. At least 20 sentences per condition and an initial signal-to-noise ratio below the real *SRT50n* were required to ensure sufficient reliability. In CI users, bias and SD became unacceptably large for a maximum speech intelligibility score in quiet below 70%. In conclusion, SA algorithms with word scoring in adaptive speech-in-noise tests are applicable to various listeners, from CI users to NH listeners. In CI users, they lead to efficient estimation of the *SRT50n* as long as speech intelligibility in quiet is greater than 70%. SA procedures can be considered as a valid, more efficient, and alternative to clinical adaptive procedures currently used in CI users.

Many cochlear-implant (CI) recipients and hearing-impaired people experience difficulties with understanding speech in a noisy environment. To characterize a subjects' ability to listen in noise, speech-in-noise tests have been developed in many languages. For clinical use of a test, it is important that the test is accurate in the sense that the test should have a small test–retest variance and bias. With an accurate test, a clinician is able to measure differences between amplification and signal processing settings. Furthermore, the test should be efficient to be applicable in a busy clinic and to prevent fatigue. Efficiency here means that a sufficient accuracy is reached within a limited number of trials.

A frequently used measure of speech perception in noise is the speech reception threshold in noise (*SRT50n*), defined by the signal-to-noise ratio (SNR) that yields an average response of 50% correctly recognized items over a number of trials (Plomp & Mimpen, 1979). This *SRT50n* can be measured with an adaptive procedure that varies the SNR based on previous responses of the listener to track the 50% score. The SNR and the percent correct score are related by a psychometric curve, which is often referred to as the intelligibility function. The slope of this curve is steepest

Department of Otorhinolaryngology and Head and Neck Surgery, Erasmus Medical Center, Rotterdam, the Netherlands

Corresponding Author:
Gertjan Dingemanse, Department of Otorhinolaryngology and Head and Neck Surgery, Erasmus Medical Center, Room NT-310, P.O. Box 2040, 3000 CA, Rotterdam, the Netherlands.
Email: g.dingemanse@erasmusmc.nl

around the 50% correct score in normal-hearing (NH) listeners. The adaptive procedure keeps the trials in this steep part of the curve and avoids potential floor and ceiling effects. In general, tests of sentence recognition in steady-state speech-spectrum noise have intelligibility functions with steep slopes, giving the advantage that the *SRT50n* estimate is accurate, because the test–retest variance is inversely related to the slope (e.g., Kollmeier et al., 2015). The slope of the intelligibility function is often increased by optimizing the homogeneity of the sentences with respect to their *SRT50n* and slope.

For CI users, speech-in-noise tests may not be optimally designed. First, the just-mentioned optimization of the homogeneity of the sentences is usually done in a group of NH listeners, and it is unknown whether this homogeneity also applies to CI users. Second, the slope is often less steep in CI recipients. Dingemanse and Goedegebure (2015) found an average slope of 6.4%/dB around 50% for CI recipients, which is much lower than the typical slope of 10%/dB to 15%/dB obtained with NH listeners (e.g., Versfeld et al., 2000). However, the step sizes used in adaptive speech tests are often the same in CI recipients as in NH listeners (e.g., Chan et al., 2008; Dawson et al., 2011; Zhang et al., 2010), which may result in different step size to slope ratios for CI recipients compared with NH listeners. This can reduce the accuracy of the adaptive procedure. Third, the maximum proportion correct score (measured in quiet) is lowered and may range from 1 to 0.1 (e.g., Gifford et al., 2008), making the proportion correct score of 0.5 no longer the point with the steepest slope. Consequently, the accuracy of the *SRT50n* measure may be insufficient for CI listeners, or an adaptive estimation of the *SRT50n* is not even feasible if the maximum proportion correct score of a CI listener approaches 0.5. Given these concerns, there is a need to address the accuracy of *SRT50n* measures in CI listeners and to explore if *SRT50n* measurements need special procedures in CI listeners to enhance accuracy.

Several researchers have attempted to modify the simple up-down procedure for use in CI recipients because of their reduced speech intelligibility. The Hearing in Noise Test procedure was modified by allowing one or more errors in repeating a sentence (Chan et al., 2008) or allowing a maximum error of 20%, 40%, or 60% (Wong & Keung, 2013). Wong and Keung showed that adaptive procedures based on these criteria could be used in a greater percentage of CI users. These modifications of the scoring may improve the accuracy because of the increase in maximum proportion correct score and the slope at *SRT50n*.

Another well-known option to enhance the accuracy of the *SRT50n* estimate is to score the correctly repeated sentence elements (often words, so-called word scoring; Brand & Kollmeier, 2002; Terband & Drullman, 2008).

The test–retest reliability is inversely proportional to the square root of the number of sentences and for word scoring also to the number of statistically independent elements per sentence. The effective number of statistically independent elements in a sentence is typically around two words per sentence. This is less than the number of words in the sentence because the words in a sentence are related by the contextual information of the sentence (Boothroyd & Nittrouer, 1988). In CI users having a lowered maximum proportion correct score, word scoring is a good option, because this type of scoring can still be used, while sentence scoring is not feasible.

If word scoring is used, an adaptive procedure has to prescribe how the step size depends on the proportion of correct words. Hagerman and Kinnefors (1995) described such a procedure. They used small step sizes if only some of the words were recognized and larger steps if all words or none of the words were recognized. Brand and Kollmeier (2002) proposed a generalization of the Hagerman and Kinnefors procedure based on the difference between the proportion of correct words in the previous trial and the target proportion correct. This difference was divided by the slope of the intelligibility function and scaled by a scaling function that governed the step size sequence. A concern with this adaptive procedure is that the optimal step size is related to the slope of the intelligibility curve, which is most often unknown and can vary considerably in CI users and hearing-impaired listeners.

The accuracy of an *SRT50n* estimate also depends on the adaptive procedures themselves and the way in which the *SRT50n* is calculated. Often, adaptive procedures use a fixed step size to govern SNR placement and the average SNR over the trials as the *SRT50n* estimate (Nilsson et al., 1994; Plomp & Mimpen, 1979). These simple up-down procedures are nonparametric. Several researchers used a parametric maximum-likelihood estimation of the *SRT50n* and the slope, with the aim of improving accuracy (Brand & Kollmeier, 2002; Versfeld et al., 2000). However, Versfeld et al. showed that maximum-likelihood estimates were not systematically different from an estimate based on the average of the last 10 sentences of the nonparametric simple up-down procedure. Others have proposed Bayesian methods to estimate the parameters of the psychometric function (King-Smith & Rose, 1997; Kontsevich & Tyler, 1999). Such methods can also be used to control SNR placement (e.g., Doire et al., 2017; Shen & Richards, 2012). In general, both maximum-likelihood estimation and Bayesian estimation require some prior knowledge of the intelligibility function. Most studies have assumed the maximum proportion correct near 1 and did not test the performance of an estimation method for a lower maximum proportion correct score (but cf. Green, 1995). Shen and Richards (2012) proposed

a method that includes an estimation of the maximum proportion correct. A disadvantage of their method is that all parameters of the psychometric function must be estimated concurrently, which requires a larger number of trials at well-distributed SNRs. In contrast, nonparametric methods only assume a monotonic increasing intelligibility function (cf. Robbins & Monro, 1951) and are able to estimate the *SRT50n* as the only parameter. Although some prior knowledge of the mean and slope may help to optimize nonparametric adaptive procedures, this knowledge is not a fundamental requirement. Furthermore, nonparametric methods are easier in concept and calculation.

The nonparametric adaptive procedures are in fact stochastic approximation (SA) methods that try to approximate the *SRT50n* based on scores from earlier trials, which are stochastic in nature. SA algorithms were originally developed to find the roots of a function if only noisy observations are available (Robbins & Monro, 1951). In the context of this study, it means to find the root of the function $f(SNR) - 0.5$, in which f is the intelligibility function. Nowadays, there is a large body of literature on SA describing a variety of recursive SA algorithms with different step size sequences (for an overview, see Kushner & Yin, 2003).

SA algorithms often have step size sequences that decrease with increasing trial number n. The rationale is that the estimation of the root (or target proportion correct) is more accurate if the step size decreases during the recursive approximation (Kushner & Yin, 2003). Decreasing step size sequences have also sometimes been used for speech-in-noise measurements (Brand & Kollmeier, 2002; Keidser et al., 2013).

A concern of using a decreasing step size sequence in speech tests is that it makes an adaptive threshold estimation algorithm more prone to bias due to nonstationary behavior of the listener, such as lapses in attention. Fatigue can also occur, although Dingemanse and Goedegebure (2015) have found no effect of fatigue in a typical experiment with CI users. A second concern regarding the use of decreasing step sizes is that there is a risk of bias if the SNR of the first trial is relatively far from the real *SRT50n*. So, when using SA algorithms with decreasing step sizes, consideration should be given to possible effects of nonstationary behavior of the listener and the selection of the initial SNRs.

The aim of this study is to find an efficient SA algorithm for *SRT50n* estimation in CI users, using word scoring, and taking into account intelligibility functions with less steep slopes and a lower maximum intelligibility score in quiet.

The research questions are as follows:

1. Is there an SA algorithm based on word scoring that provides a more efficient estimate of the *SRT50n* than clinically used procedures in CI users?

2. What are the conditions for reliable use of adaptive measurements of *SRT50n* in CI users, with respect to the speech intelligibility score in quiet and the initial SNR?

To answer these questions, we selected several SA algorithms from the literature. We used Monte Carlo simulations to investigate the efficiency and accuracy of the SA algorithms. The main outcome measures were the standard deviation (SD) and the bias of the estimated *SRT50n*. Simulations with NH subjects were included to get insight into possible differences in optimal algorithms or parameters between CI recipients and NH listeners.

Materials and Methods

SA Algorithms

To find the root of a function $f(SNR) - P_t$, with P_t the target proportion correct, SA algorithms use an adaptive up-down procedure of the form:

$$x_{n+1} = x_n + a_n(P_t - y_n) \tag{1}$$

where x_n is the stimulus value (the SNR) of the nth trial, y_n the proportion of correctly recognized words as a noisy measurement of the value $f(x_n)$, P_t the target proportion correct, and a_n the step size parameter of the nth trial. Robbins and Monro (1951) proved that a decreasing step size sequence of $a_n = b/n$ implies convergence of x_n to x_t with $f(x_t) = P_t$, where b is the step size constant, and f a monotonically increasing function. In the literature on SA, many other step size sequences and their convergence are described, and even other recursive formulas have been proposed (Kushner & Yin, 2003).

For our purpose, we need SA algorithms that have the following properties: (a) a good small-sample convergence because sentence lists have a relatively small number of trials (10–30 sentences) for reasons of test efficiency; (b) good rejection of the noise in the y_n because the variance of the noise in y_n is large; (c) insensitivity to badly chosen initial values or large deviations of y_n from P_t early in the procedure to prevent bias; and (d) tolerance with respect to some nonstationarity in the intelligibility function due to nonstationary behavior of the participants, such as varying attention. Note that these four requirements describe different aspects but are not independent of each other. In general, smaller step sizes are better for noise rejection, and larger step sizes lead to faster forgetting of initial conditions.

In the SA literature, four algorithms were found that may meet the aforementioned criteria. The first algorithm is the accelerated SA (Kesten, 1958). Kesten

proved that the convergence of the SA sequence can be accelerated compared with the original form (Equation 1) if the step size decreases on reversals of the direction of the iterates.

$$x_{n+1} = x_n + a_{n_{rev}}(P_t - y_n), \quad a_{n_{rev}} = \frac{b}{n_{rev} + 1} \quad (2)$$

where n_{rev} is the number of reversals. The last iterate x_{n+1} is the estimate of the x_t for which $f(x_t) = P_t$. The accelerated SA has good small-sample convergence. We need to determine the optimal value of b for speech tests.

A second algorithm is the averaged SA with decreasing step size (dss) sequence (averaged dss SA). It uses the original algorithm of Equation 1 together with averaging of the iterates:

$$x_{n+1} = x_n + a_n(P_t - y_n), \quad a_n = \frac{b}{n^\alpha} \text{ and } \bar{x}_n = \frac{1}{n - n_e} \sum_{i=1+n_e}^{n} x_i \quad (3)$$

with step size decrease rates α. Because x has to converge to the target, it is likely that the first trials are not close to the target. Therefore, the first n_e trials may be left out of the average. The result of the average $\bar{x}_{n+1}$ gives the estimate of x_t. In the SA literature, this is known as Polyak–Ruppert averaging (Polyak, 1990; Polyak & Juditsky, 1992; Ruppert, 1988). It was shown by Polyak and Juditsky (1992) that this average is preferable if the step size sequence $[a_n]$ goes to zero slower than order $1/n$. The idea is that relative large step sizes $[a_n]$ lead to faster forgetting of initial conditions, while use of the average reduces noise. In the original form, $n_e = 0$, but it is also possible to introduce exclusion of the initial values with $n_e > 0$. For this algorithm, we need to determine the optimal step size sequence parameters b, α, and n_e.

A third option is the use of a not decreasing step size (ndss) sequence together with averaging (averaged ndss SA). In fact, this is the Polyak–Ruppert averaging from Equation 3 with $\alpha = 0$ and $a_n = b$. This option was used in speech recognition tests by Hagerman and Kinnefors (1995). They proposed a procedure with $P_t = 0.4$ and $a_n = b = 5$ for five-word sentences. If applied to six-word sentences, as in this study, the procedure is implemented by choosing $P_t = 0.5$ and $a_n = b = 6$.

A fourth algorithm that may be suitable to use with a speech test is the so-called smoothed SA that was first described by Bather (1989) and was further considered by Schwabe (1994; Schwabe & Walk, 1996). In this algorithm, the average of both the iterates x_n and the noisy observations y_n are used in the recursive equation:

$$x_{n+1} = \bar{x}_n + n \, a_n \, (P_t - \bar{y}_n) \quad (4)$$

where

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i , \quad \bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ and } a_n = \frac{b}{n^\alpha} \quad (5)$$

The average of the iterates $\bar{x}_{n+1}$ is the estimate of x_t, also with the possibility to exclude the first n_e trials. Schwabe and Walk (1996) showed that for step sizes with $1/2 < \alpha < 1$, the influence of inappropriate starting points decays faster than in Polyak–Ruppert averaging.

Simulation Model of a Listener

To be able to test the accuracy of the proposed SA algorithms with Monte Carlo simulations, we have made a simulation model of speech recognition that generates a listener's response for a given SNR.

The first element of the listener model is an intelligibility function that describes the average proportion correct words in a sentence as a function of the SNR. The intelligibility function was modeled as

$$p(SNR) = \frac{(1 - \lambda) \, p_{max}}{1 + \exp(4s(SRT_m - SNR))} \quad (6)$$

with p the proportion of correctly recognized words in a sentence, λ the lapse rate, p_{max} the proportion correct in quiet, SRT_m the x where $p(x)$ is half $(1–\lambda) \cdot p_{max}$, and s the nominal slope (the slope of p at SRT_m is $(1–\lambda) \cdot p_{max} \cdot s$). For higher p, lapses may occur due to moments of inattentiveness, and for low p, there may be some lapsing because the listener gives up (Bronkhorst et al., 1993).

The intelligibility function was fitted to the data of a group of 20 CI users from a study of Dingemanse and Goedegebure (2015). In that study, speech intelligibility in noise was measured at three SNRs, with three corresponding performance levels: adaptively estimated SRTs at 50% and 70% words correct and performance level at a fixed SNR of SRT50% + 11 dB. The performance was measured with and without activation of a noise reduction algorithm. Furthermore, speech intelligibility in quiet was measured. For each of the participants, the intelligibility function was fitted to all the data because the noise reduction algorithm had no measurable effect on the speech performance. Table 1 shows mean, SD, and range of the group for the different parameters of the intelligibility function. Only relatively high-performing CI users were included. SRT_m and s were not significantly correlated.

The intelligibility function was also fitted to the data of a reference group of 16 NH subjects with a mean age of 22 years, described by Dingemanse and Goedegebure (2019). In that study, the $SRT50n$ was adaptively

Table 1. Values of the Parameters of the Intelligibility Function (see Text at Equation 6) for a Group of CI Recipients and a Group of NH Listeners.

	CI group				NH group			
	M	*Mdn*	*SD*	Range	*M*	*Mdn*	*SD*	Range
SRT_m (dB)	3.7	3.4	2.7	−1.0–10.7				
SRT50n (dB)	4.2	3.4	3.3	−1.0–12.7	−5.5	−5.5	0.6	−6.6 − −4.6
s (pc/dB)	0.067	0.065	0.021	0.029–0.125				
s50 (pc/dB)	0.064	0.064	0.021	0.026–0.122	0.151	0.146	0.025	0.116–0.192
p_{max} (pc)	0.947	0.965	0.062	0.740–1.0	1.0	1.0	0	1.0–1.0

Note. The mean, median, SD, and range are given. For the NH group, the SRT_m and the SRT50n are the same, and s and s50 are the same.
pc = proportion correct; SRT50n = the speech reception threshold at a proportion of correctly recognized words of 0.5; s50 = the slope at the 0.5 point; CI = cochlear implant; NH = normal hearing; SD = standard deviation.

measured using word scoring and the ndss SA algorithm with $b = 4$, along with the proportion of correct words at four SNRs around the individual *SRT50n*. The intelligibility function was fitted to the performance at these four SNRs, assuming that $\lambda \approx 0$. Table 1 shows the parameter values found. In both studies, Vrije Universiteit (VU) sentences (2 lists of 13 sentences for each condition) and steady-state speech-spectrum noise were used (Versfeld et al., 2000).

In practice, variation in intelligibility from trial to trial occurs due to variability in the SRT and slope of sentences, differences between listeners, and variability in listening effort and attention. We modeled variability in sentences by adding a normal distribution of SRT_m values with a small SD $SD\_SRT_m = 0.5$ dB and a normal distribution of variation is slopes with $SD\_slope = 0.01$. These values were in accordance with Versfeld et al. (2000). To incorporate differences between subjects, variation of *SRT50n* between subjects was modeled as a normal distribution with an SD of 1 dB for the NH group (based on Versfeld et al., 2000) and 3 dB for the CI group (based on Table 1). The variation in slope between listeners was varied according to a normal distribution with an SD of 0.02, according to Table 1. To account for variability in attention, the lapse rate (λ in Equation 6) was set to 0.02 independent of the proportion correct score. This means that in 2% of the trials the listener is not attentive.

The second element of the listener model models the response of a listener in a trial. For this element, a multinomial distribution is used, giving the probabilities that k out of l words ($k = 0, \ldots, l$) of a sentence were correctly recognized as a function of the average proportion correct word score. The multinomial distribution was obtained from a model of Bronkhorst et al. (1993, 2002) for context effects in speech recognition. This model gives predictions of the probabilities $p_{w,k}$ that k elements ($k = 0, \ldots, l$) of wholes containing l elements are recognized. These probabilities $p_{w,k}$ are a function of a set of context parameters c_i ($i = 1, \ldots, l$) and the

recognition probabilities of the elements if presented in isolation (no context) $p_{i,nc}$.

$$p_{w,k} = F(c_i, \; p_{i,nc}), \quad 0 \leq c_i \leq 1, \quad i = 1, \ldots, l \quad (7)$$

The context parameters c_i give the probabilities of correctly guessing a missing element given that i of the l elements were missed. They quantify the amount of contextual information used by the listener. The maximum value of 1 means that a missing element is available from context information without uncertainty. The minimum value is the guessing rate if the whole contains no context information. For details of the model, we refer to Bronkhorst et al. (1993, 2002). From the array of $p_{w,k}$ values, we can calculate the average proportion of correctly recognized words in sentences:

$$p_e = p_{w,l} + \frac{(l-1)}{l} p_{w,l-1} + \frac{(l-2)}{l} p_{w,l-2} + \ldots + \frac{1}{l} p_{w,1}$$

(8)

This model was fitted to speech recognition data of a group of CI users and a group of NH listeners from the study of Dingemanse and Goedegebure (2019), resulting in a set context parameters for each group (see their Figure 4). In the study of Dingemanse and Goedegebure, VU sentences (Versfeld et al., 2000) were used as speech material in both groups.

Figure 1 shows in the left panel the probabilities $p_{w,k}$ as a function of p_e for the CI group. For example, at the 50% correct point of the intelligibility function ($p_e = 0.5$) in 25% of the trials, the whole sentence is recognized ($k = 6$), but in another 25%, no words are recognized ($k = 0$); this is illustrated in the right panel of Figure 1.

In the Monte Carlo simulations, the response of a listener in a trial was obtained following the next steps: First, the average word recognition probability was calculated from the intelligibility function (Equation 6) for the SNR of the trial, resulting in value p_x. Next, a
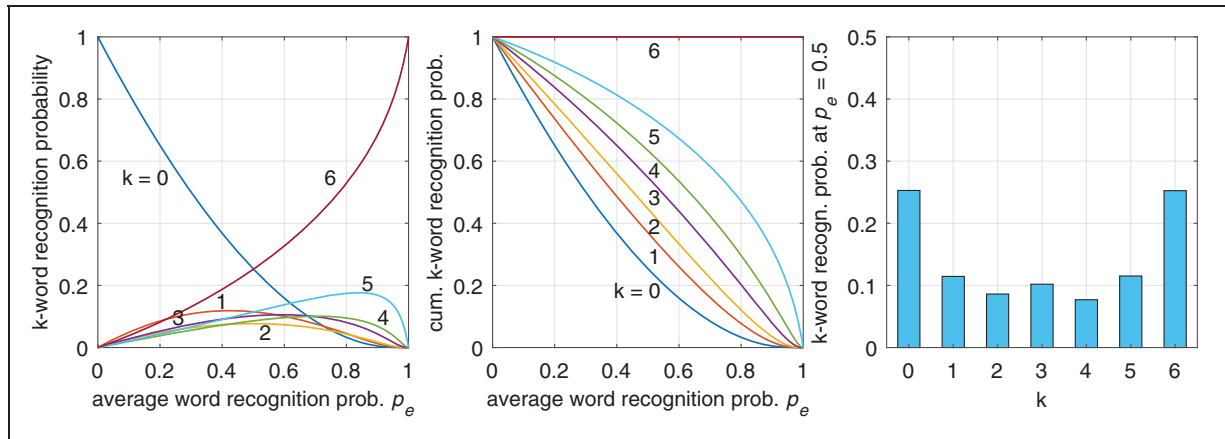
Figure 1. Left panel: Probabilities to recognize k words of a sentence correctly as a function of the average proportion of correctly recognized words p_e. Center panel: Cumulative probabilities to correctly recognize k words or less as a function of p_e. Right panel: Example of the multinomial distribution for an average word score of $p_e = 0.5$ that gives the probability to recognize k words from a sentence.

random number from a continuous uniform distribution with a minimum value of 0 and a maximum value of 1 was taken, giving value p_y. Third, point (p_x, p_y) was compared with the cumulative probabilities shown in the center panel of Figure 1. For example, the point of $p_x = 0.5$ and $p_y = 0.7$ fell in the area of $k = 5$. That is, five out of six words were correctly recognized in this trial.

We added some variation in the context parameters using a normal distribution with an SD of 0.01 for c_1 to 0.016 for c_5 to simulate differences between listeners (Dingemanse & Goedegebure, 2019).

Validation of the Simulation Model

The validity of the model for the description of averaged speech recognition scores has already been demonstrated by Bronkhorst et al. (1993, 2002). To verify if the model not only describes speech recognition on average but also produces reliable word scores for single trials in adaptive procedures, we used the within-staircase SD as a measure to compare simulation outcomes with experimental data. The within-staircase SD shows whether the simulation model produces realistic variations within a staircase. As the model parameters were tuned to the CI group of Dingemanse and Goedegebure (2015), the model should produce the same within stair-cases as found in the experimental data. The *SRT50n* staircases were measured in two conditions in Dingemanse and Goedegebure (2015). The mean within-staircase SD was calculated as the root mean square (RMS) of the individual within-staircase SDs from the two conditions and resulted in a value of 2.0 dB. The adaptive procedure used was the averaged ndss SA, with $b = 4$. Simulations with this procedure

resulted in a within-staircase SD of 2.1 dB. This corresponds very well with the experimental value of 2.0 dB.

When parameters of the NH group were applied, a within-staircase SD of 1.5 dB was found, which is in good agreement with the 1.4 dB found from the *SRT50n* measure in Dingemanse and Goedegebure (2019). From the same study, a within-staircase SD of 1.9 dB for sentence scoring combined with a fixed step size of 2 dB and 13 trials was available. The within-staircase SD of the simulation of this condition was also 1.9 dB.

Versfeld et al. (2000) reported that the within-subjects SD of the *SRT50n* was 1.1 dB for sentence scoring and an adaptive up-down procedure with a 2 dB step size. A simulation of this condition resulted in a within-subjects SD of 1.1 dB.

These results confirmed the validity of the used listener model for use in simulations of adaptive procedures.

Calculation of Reference SDs at SRT50n

The listener model was used to generate 4,000 responses based on word scoring at an SNR of *SRT50n*. The SD of these responses was calculated and served as a reference measure of the variability in proportion correct speech recognition at the *SRT50n* due to the stochastic nature of the speech recognition process. Table 2 presents the reference SDs of the simulations at a fixed SNR of *SRT50n*. The calculated SD was divided by the slope of the intelligibility function at the *SRT50n* point to obtain a reference SD of the *SRT50n* measure. The SDs of the *SRT50n* estimates of the SA algorithms were compared with these reference SDs to get a

Table 2. Reference Standard Deviations of Proportion Correct Words From Sentences P_t and *SRT50n* Values, Resulting From Simulations of CI and NH Listeners at a Fixed SNR of *SRT50n*.

Sentence list length	CI group		NH group	
	SD Pt	SD SRT50n	SD Pt	SD SRT50n
13	0.137	2.33	0.121	0.824
20	0.104	1.77	0.091	0.616
26	0.089	1.52	0.078	0.528

Note. SRT50n = speech reception threshold in noise; CI = cochlear implant; NH = normal hearing; SD = standard deviation.

measure of the variability introduced by the SA algorithms itself.

In the simulation model, small variations in *SRT50n* and slope between sentences and between subjects were included, as mentioned in the model description. By comparing the simulation results with and without applying variations, it turned out that the effect of the variations in model parameters was a 4% to 6% increase of the SDs in CI users and a 0.5% to 1.3% increase in NH users.

The SDs of the P_t estimates in the CI group were slightly greater than the SDs of the NH group due to the fact that the model for CI users had higher values for the context parameters. The SDs of *SRT50n* are higher in CI users because the slope of the intelligibility function is less steep. SDs decreased approximately with the square root of the list length, bearing in mind that the first four sentences were excluded in the calculations for all list lengths.

Simulation Procedures

In the simulations, we used a slope of $0.15\,dB^{-1}$ for NH users and half that value for the CI group (Equation 6). The parameter p_{max} was set to 1 for NH listeners. For relatively high-performing CI users, the value was 0.95 according to Table 1. To represent a broader range of performance values between 0.6 and 1, p_{max} was set to 0.8 for CI users. The initial SNR (the SNR of the first trial) relative to the mean *SRT50n* was taken from a normal distribution with mean $= -3\,dB$ (NH) or $-6\,dB$ (CI) and $SD = 1\,dB$ (NH) or $3\,dB$ (CI). The first trial was repeated at increasing SNRs ($+2\,dB$) until at least half of the words were recognized correctly or the sentence was three times repeated.

In the simulations, independent streams of random numbers were generated for each variable for which a probability distribution was defined. For each condition, 2,000 simulations of staircases were generated, and each staircase consisted of 26 trials. For each simulation, the

SRT50n estimate was the average or the end value of the staircase, depending of the SA algorithm. For each condition, three outcome measures were calculated: the SD and bias of *SRT50n*, and the within-staircase SD calculated as the RMS average of the 2,000 SDs of the SNRs within each staircase. We calculated the three outcome measures for sentence list lengths of 13, 20, and 26 sentences, as the minimum list length is 13 sentences for the speech material used in the model. A length of 26 sentences (two lists) is around a maximum length that can be used in clinical settings, in our opinion. A length of 20 sentences is included because this list length is used in other speech material (e.g., Soli & Wong, 2008), and it is in the middle of the clinically feasible range of the number of sentences to be used. All simulations and analyzes were performed with MATLAB (9.6.0, The MatWorks Inc., Natick, Massachusetts, USA).

Finding Optimal Parameters for SA Algorithms

To find optimal values of the parameters in the SA algorithms, simulations were performed while varying the relevant parameters. The step size constant *b* was varied from 2 to 14 dB in steps of 2 dB for the CI group and from 1 to 7 dB in steps of 1 dB for the NH group. Because the maximum of $(y_n - p_t)$ in the Equations 1 to 4 is 0.5, $b = 4$ corresponded to the often used step size of 2 dB. For the averaged dss SA and the smoothed SA, optimal parameters were determined by simulations for step size decrease rates α from 0.1 to 0.5 with a step of 0.1 for the averaged dss SA and from 0.5 to 1 (step 0.1) for the smoothed SA. For the averaging SA algorithms, the number of excluded trials n_e was 4, 6, or 8 trials.

To find the best parameter set of *b*, α, and n_e, we looked for minimum SD and bias of *SRT50n* for each combination of *b*, α, and n_e. However, the minima of SD and bias were often not reached at the same parameter values. We regarded a minimum SD as the most important criterion (i.e., for test–retest purposes), but we did not allow differences in intelligibility greater than 5% due to bias because that may become a clinically relevant difference. Based on this criterion, the mean bias should be $\leq 0.85\,dB$ in the CI group and $\leq 0.33\,dB$ in the NH group. The parameter set that produced the smallest SD of *SRT50n* within these bias criteria was chosen as the optimal parameter set of *b*, α, and n_e. The optimization was done for each of the three list lengths.

Simulations With the Optimal SA Algorithms and Clinical Procedures

In the simulations, we also included some clinically used procedures. First, sentence scoring with a fixed step size of 2 dB was included (Nilsson et al., 1994; Plomp

& Mimpen, 1979). Second, a procedure of modified sentence scoring was added, allowing 2 errors per sentence (66.67%) such as in Chan et al. (2008) and Wong and Keung (2013). In this procedure, the SNR was varied adaptively as in Chan et al., that is, in 5 dB steps for the first four sentences and in 3 dB steps for the remaining sentences of the list for the CI group. For the NH group, the steps were 4 dB for the first four sentences and 2 dB for the remaining trials as in the Hearing in Noise Test procedure (Soli & Wong, 2008). Because the psychometric curve of Equation 6 applies to word scoring, we calculated the change of the psychometric curve from the context model (Equations 7 and 8) for sentence scoring and modified sentence scoring. Figure 2 shows the resulting curves.

Furthermore, we included a third clinically used procedure based on word scoring: the procedure of Brand and Kollmeier (2002). They proposed the following formula:

$$x_{n+1} = x_n + a_{n_{rev}}(P_t - y_n), \quad a_{n_{rev}} = \frac{1.5 \cdot 1.41^{- n_{rev}}}{slope} \quad (9)$$

We used $p_{max} \cdot s$ as slope value. Brand and Kollmeier used a maximum-likelihood estimate of the $SRT50n$, but because only nonparametric methods are investigated in this study, the last iterate x_{n+1} was used as an estimate of the threshold x_t. Henceforth, this procedure will be referred as the npBK SA procedure.
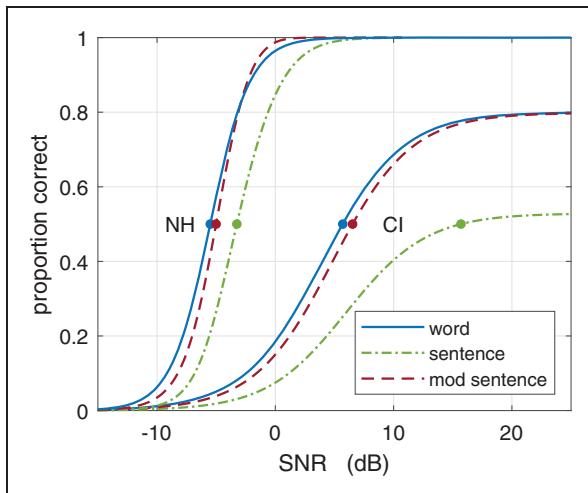

Figure 2. Intelligibility Functions of Correctly Recognized Words From Sentences, Sentence Scoring, and Modified Sentence Scoring. The three leftmost curves represent the functions of the NH group, and the three rightmost curves represent the functions of the CI group. Dots show the target proportion correct of 0.5. NH = normal hearing; CI = cochlear implant; SNR = signal-to-noise ratio.

We performed simulations with each optimized SA algorithm and the clinical procedures to investigate how their accuracy depends on the relative initial SNR by varying this SNR from –8 dB to +8 dB relative to the real $SRT50n$ value. In these simulations, the first trial was not repeated. In addition, we examined the effect of the maximum intelligibility in quiet. The parameter p_{max} was varied in five steps from 0.6 to 1 for each optimized algorithm, and the relative initial SNR was taken from a normal distribution, as described earlier.

Results

Simulations With SA Algorithms to Find Optimal Parameters

Based on all simulations, we selected optimal parameters for each SA algorithm for both listener groups according to the criteria given in the Methods section. Exclusion of the first four trials ($n_e = 4$) in the averaging resulted in the smallest SD and bias values of $SRT50n$ for all list lengths, compared with 6 or 8 ignored trials, although differences were small (between 0 and 0.15 dB). Therefore, only results for $n_e = 4$ were presented throughout the Results section.

For the smoothed SA, we found that the last iterate was a better estimate for $SRT50n$ with smaller SDs than the average of the iterates. So this end value was used instead of the average value.

Regarding the step size decrease rate α, it was found that a midrange value together with a moderate initial step size b resulted into the smallest SD and bias in CI users. A small initial step size and a large decrease rate resulted in a large bias. A large initial step size and a large decrease rate resulted in lower SD and bias, but even lower values were found for a moderate decrease rate and initial step size. Table 3 shows the optimal parameters and the SD and bias that were obtained with these parameters. The optimal step size decrease rate α was the same for CI and NH listeners, but the step size constant b was larger for the CI group. In CI users, the parameters given in Table 3 resulted in a bias smaller than the criterion value of 0.85 dB in the range of –8 to +4 dB for a staircase length of 26 sentences. For relative initial SNRs > 4 dB, the bias exceeded the criterion value for any set of parameter values. For a staircase length of 20 sentences, the bias exceeded the criterion value for a relative initial SNR > 3 dB. A list length of 13 sentences resulted in relatively high SDs and/or large bias (see also Figure 3) and was therefore not suitable.

Figure 3 shows the effect of the step size constant b on the SDs and biases of $SRT50n$ for the different SA algorithms (with optimal α value). The panels on top of the figure show the results for the CI group, and the bottom

Table 3. Optimal Values for the Step Size Constant *b* and the Step Size Decrease Rate α for the Accelerated SA Algorithm, the Averaged SA Algorithm With Decreasing Step Size (dss) or Not Decreasing Step Size (ndss), and the Smoothed SA Algorithm if Applied in CI Recipients and in NH Listeners.

	CI group				NH group			
SA algorithm	*b*	α	SD	*Bias*	*b*	α	SD	*Bias*
Accelerated SA	6	–	1.77	−0.40	4	–	0.55	−0.06
Averaged dss SA	6	0.3	1.65	−0.23	5	0.3	0.55	−0.02
Averaged ndss SA	4	–	1.71	−0.02	4	–	0.58	0.01
Smoothed SA	6	0.7	1.71	−0.30	4	0.7	0.55	−0.06

Note. For each optimized SA algorithm, the SD and bias of the *SRT50n* estimates are provided. SA = stochastic approximation; CI = cochlear implant; NH = normal hearing; SD = standard deviation.
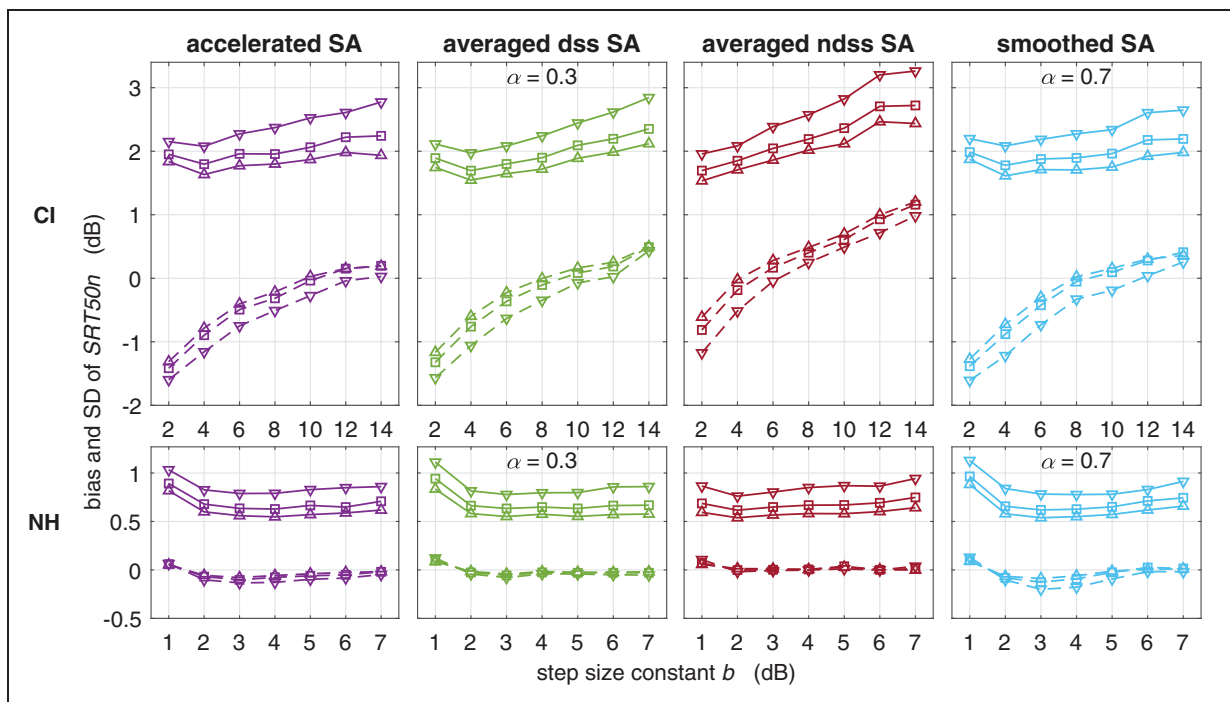

Figure 3. Estimated Values of SD (Solid Lines) and Bias (Dashed Lines) of *SRT50n* as a Function of the Step Size Constant *b* From Simulations With the Different SA Algorithms. The upper row of panels shows the results of the CI group, and the second row shows the results of the NH group. Downward-pointing triangles: 13 sentences, squares: 20 sentences, and upward-pointing triangles: 26 sentences. *SRT50n* = speech reception threshold in noise; CI = cochlear implant; NH = normal hearing; SD = standard deviation; SA = stochastic approximation; dss = decreasing step size; ndss = not decreasing step size.

panels show the results for the NH group. We observed that the SD of *SRT50n* was much greater in CI recipients than in NH listeners for all SA algorithms. In CI users, the SD was smallest for *b* = 4, except for the averaged ndss SA that had the smallest SD for *b* = 2. But for these *b* values, too much negative bias was found. Therefore, *b* = 6 (4 for the averaged ndss SA) was found to be optimal. In the NH group, the SDs of *SRT50n* were small and almost independent of *b*, indicating that the step size constant is not critical. The bias was close to zero for all algorithms and *b* values. Using a larger number of

sentences resulted in smaller SD and bias for all conditions. It is remarkable that the different SA algorithms resulted in comparable minimum SDs.

The Within-Staircase SD

The left panel of Figure 4 shows the RMS within-staircase SD as a function of the step size factor *b* for the CI group. The RMS within-staircase SD increased for increasing *b*, as expected, but differed in size between SA algorithms. The smallest values were found for
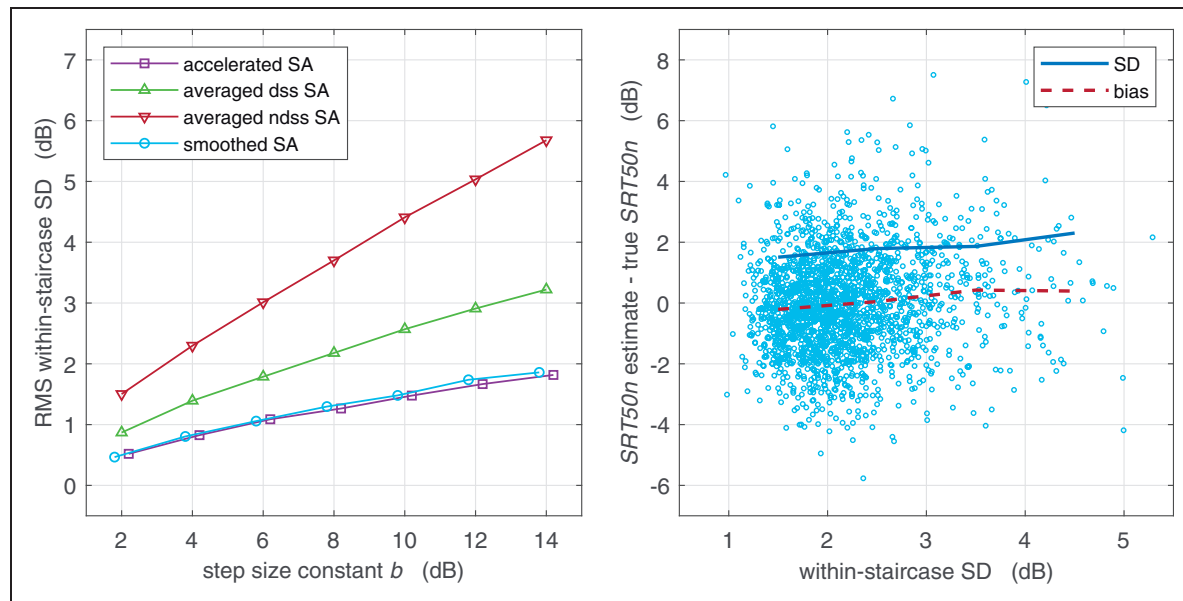
Figure 4. Left panel: RMS within-staircase SDs for the SA methods as a function of the step size constant *b* for the CI group. Each data point is calculated from 2,000 simulations. Only results for 26 trials were shown. Right panel: *SRT50n* estimates minus the true *SRT50n* plotted together with the SD and bias of the data as a function of the within-staircase SD. The data originate from 2,000 simulations with 26 trials and the averaged ndss SA algorithm, with *b* = 4.
SA = stochastic approximation; dss = decreasing step size; ndss = not decreasing step size; RMS = root mean square; SD = standard deviation; *SRT50n* = speech reception threshold in noise.

algorithms with decreasing step size. The right panel shows the *SRT50n* estimates minus the true *SRT50n* as a function of the within-staircase SD for the averaged ndss SA algorithm, with *b* = 4. The data points were grouped in bins of 1 SD width, and the mean (which is the bias) and SD were calculated for each bin and then plotted. Figure 4 shows that no clear relationship exists between the within-staircase SD and the SD or bias of the *SRT50n* estimates. This holds also for a list length of 20 sentences, for the other SA algorithms with optimized parameters, and for the NH listeners.

The Effect of the Initial SNR

Figure 5 shows the effect of the initial SNR (the SNR of the first trial relative to the true *SRT50n* of the intelligibility function) on the SD and bias of the *SRT50n* estimate. The simulations were performed with the optimal parameters given in Table 3. Figure 5 only shows results for a staircase length of 26 trials because the pattern of results for 20 trials (CI and NH) or 13 trials (NH) was very similar.

The SD and bias were very similar between the different SA algorithms over the entire SNR range. A relatively high bias was found for positive initial SNR values for the CI group. The bias was around zero, and the SDs were smallest for initial SNRs below the

true *SRT50n*. From these results, it is clear that an initial SNR below the true *SRT50n* would be preferable. In the NH group, the SD was almost independent of the initial SNR, and the bias was within ±0.2 dB.

As a validation, we compared the simulation of the ndss SA algorithm with *b* = 4 with data of the NH group from Dingemanse and Goedegebure (2019). In that study, the *SRT50n* was adaptively measured using the same algorithm and an initial relative SNR of 1 dB on average. In addition, an intelligibility function was fitted to the proportion of correct words at four fixed SNRs around the individual *SRT50n*. The SD of the individual differences between the *SRT50n* of the adaptive procedure and the *SRT50n* of the fitted intelligibility function was 0.55 dB. The SD of the simulations was 0.58 (Figure 5) and is in good agreement with the experimental SD.

The clinical algorithms had higher SDs of *SRT50n* than the SA algorithms over the entire SNR range. For the CI group, sentence scoring resulted in a high SD and a bias that showed that the adaptive procedure was hardly able to move the SNR value away from the initial SNR. This is in accordance with the almost flat intelligibility function around a proportion correct of 0.5 (see Figure 2). The modified sentence scoring resulted in a much better SD around 2.8 dB and a positive bias between 0.7 and 1.4 dB. The SD of the npBK SA algorithm is nearly as small as the SDs of the SA algorithms
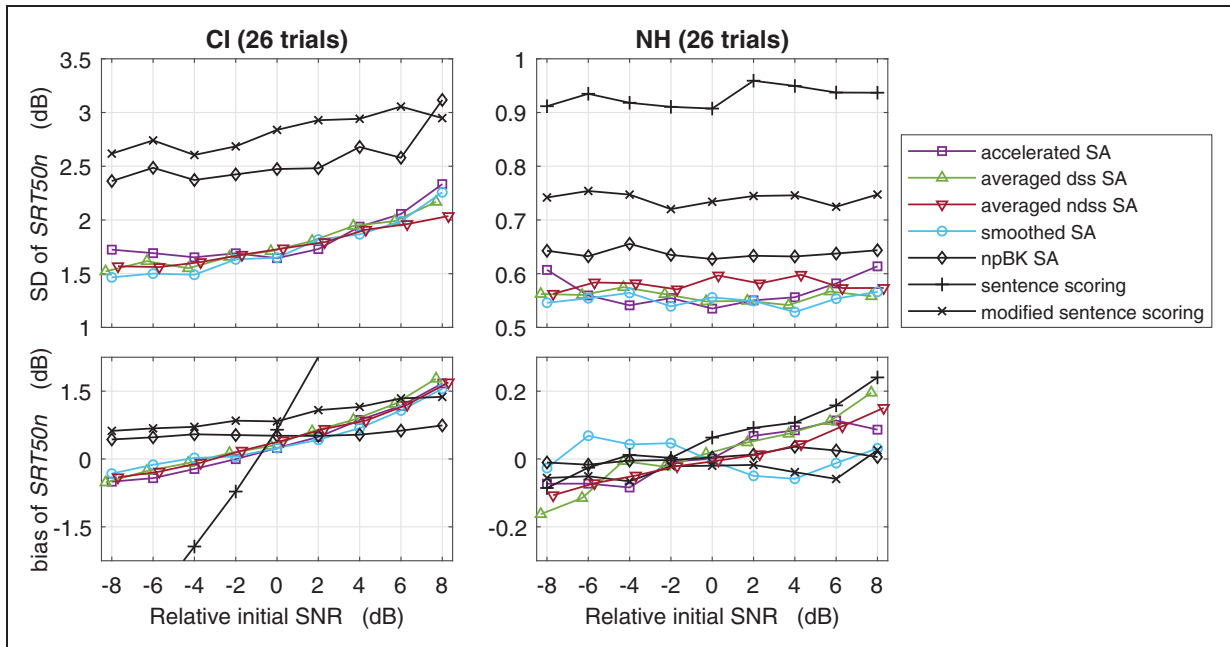
Figure 5. SD and Bias of *SRT50n* Estimates as a Function of the Initial SNR Relative to the True *SRT50n* for the SA Methods and Clinical Procedures. In the top left panel, the SD of sentence scoring is out of range. At an initial SNR of –8 dB, this SD is 4.5 dB, and it increases almost linearly to 6.5 dB at +6 and +8dB.

NH = normal hearing; CI = cochlear implant; SNR = signal-to-noise ratio; SA = stochastic approximation; dss = decreasing step size; ndss = not decreasing step size; npBK = nonparametric Brand & Kollmeijer; SD = standard deviation; *SRT50n* = speech reception threshold in noise.

in the NH group. But in the CI group, the SD is clearly greater than that of the SA algorithms, and the bias is positive.

The SA algorithms using word scoring resulted in the smallest SD and bias. For the NH group, sentence scoring resulted in an SD of 0.92 dB and only a small bias for all initial SNRs. The modified sentence scoring resulted in a smaller SD of around 0.73 dB due to the steeper slope of the intelligibility function (Figure 2), but it was still higher than the SDs of the SA algorithms that were around 0.58 dB.

The Effect of Reduced Maximum Intelligibility

The effect of p_{max} was investigated for the CI group with each of the optimal algorithms and the three clinical algorithms. Figure 6 shows that p_{max} had a large effect on the SD and bias of the *SRT50n* estimates. The SD increased for decreasing p_{max}. This effect was most apparent for sentence scoring, modified sentence scoring, and the npBK SA algorithm. For the range of p_{max} between 0.7 and 1, the SA algorithms were efficient, that is, close to the reference SD from Table 2 that serves as a theoretical minimum. At $p_{max} = 0.6$, bias values become more negative on average. Only the results for a staircase length of 26 trials were shown

because the pattern of results for 20 trials was very similar, with small bias and efficient estimation for $p_{max} \geq 0.7$.

Discussion

SA Methods Versus Clinical Procedures

The four SA algorithms proposed in this study provide more efficient estimates of the *SRT50n* than clinically used adaptive procedures in CI users, as can be observed from Figures 5 and 6. The SD estimates of the four SA algorithms were close to the reference SDs from Table 2, indicating that the SA algorithms add little variance to the *SRT50n* estimate, compared with the variability due to the stochastic nature of the speech recognition process. Even with the more shallow intelligibility functions found in CI users, the algorithms remain efficient, provided that $p_{max} \geq 0.7$ and the initial SNR is within –8 to +4 dB of the real *SRT50n*.

Several researchers recognized the inaccuracy of sentence scoring in CI users and proposed a modified sentence scoring that allows some errors per sentence (Chan et al., 2008; Wong & Keung, 2013). Indeed, the modified sentence scoring resulted in better accuracy. But the SA algorithms had both smaller SD and bias, especially
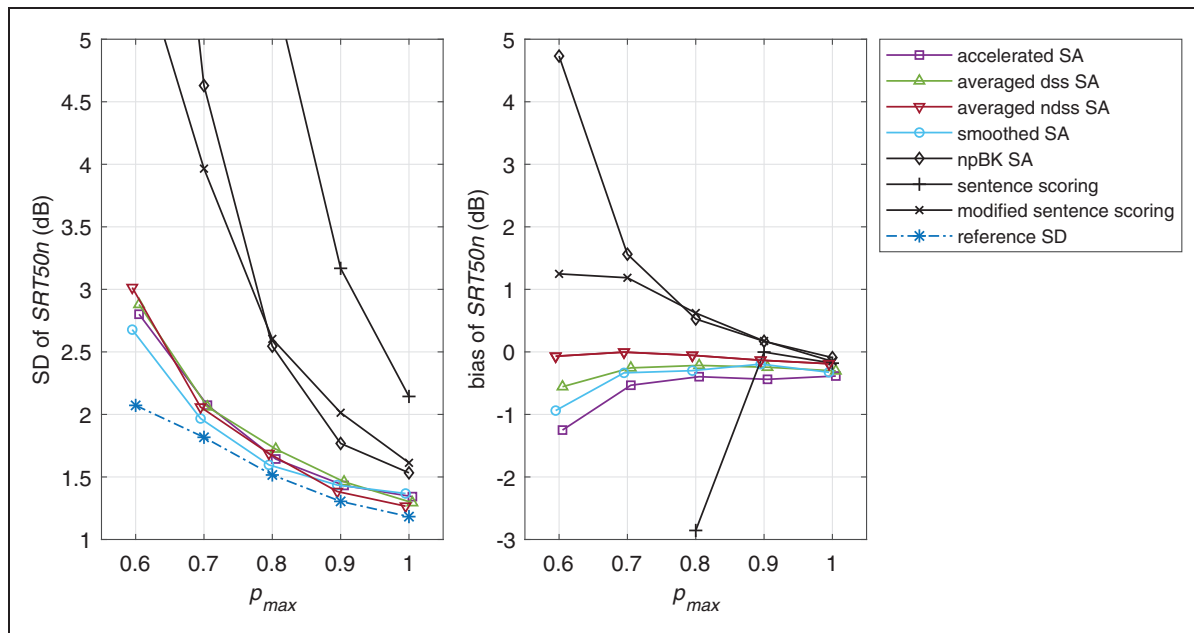
Figure 6. SD and Bias of *SRT50n* Estimates as a Function of p_{max} for the SA Methods and Clinical Procedures Applied in the CI Group. Only results of the conditions with 26 trials were shown. The dash-dotted line with asterisks gives the minimum SD based on the reference SD in Table 2 as a function of p_{max}.
SA = stochastic approximation; dss = decreasing step size; ndss = not decreasing step size; npBK = nonparametric Brand & Kollmeijer; SD = standard deviation; *SRT50n* = speech reception threshold in noise.

when p_{max} is below 1 (Figure 6). This can be explained by their use of word scoring that has a higher number of statistically independent elements per sentence, as explained in the Introduction section.

The new proposed SA algorithms also performed better than the npBK SA algorithm. The main reason is that this algorithm has relatively large steps early in the staircase and a high decrease rate. Especially in the CI group, having a lowered p_{max}, this combination resulted in a larger SD and bias. The large steps early in the staircase may result in high SNR values, were the intelligibility function is already flat. In this flat part of the function, the SNR may jump randomly up and down at high SNRs, while the step size is decreasing. As a result, the staircase ends with a large positive bias.

The four SA algorithms proposed in this study resulted in comparable SD and bias if parameters were used that were optimal for the group that was tested. There is no clear winner. It is noteworthy that a more complex SA method, such as the smoothed SA, did not result in better performance than the simpler ndss SA method. The optimal step size decrease rate α was the same in CI and NH listeners, both for the averaged dss SA and for the smoothed SA algorithm. The only difference between groups is the step size constant b, except for the averaged ndss SA algorithm, where $b = 4$ applies to both groups. The NH group and the CI group represent the extremes of the intelligibility function. The

group of people with sensorineural hearing loss, using hearing aids or not, is expected to have intelligibility functions with slopes in between the slopes of the NH group and the CI group. So, the averaged ndss SA algorithm with a step size constant of 4 is applicable to a wide range of hearing-impaired listeners. This algorithm was already used in speech recognition tests by Hagerman and Kinnefors (1995). Furthermore, it was used in several studies with CI recipients and provided highly reproducible and consistent data (cf. Dingemanse & Goedegebure, 2015, Figure 3; 2018, Figure 3; Vroegop et al., 2017).

The use of simulations gave the possibility to gain insight into the occurrence of a bias. Because the true *SRT50n* of the listener model is known, the bias can be calculated, which is impossible in real subjects with unknown *SRT50n*. In NH listeners, the bias was close to zero for all SA algorithms if initial SNRs were within −8 to +8 dB relative to *SRT50n*. If in the first trials a large step in the wrong direction is made due to the stochastic behavior of the speech recognition process, then the average proportion correct at the next SNR is much higher or lower because of the steep slope of the intelligibility function. This leads to a high chance that a reversal occurs and that is why no bias occurs. Furthermore, the intelligibility function is symmetrical in the *SRT50n* point in NH listeners, making that steps from above or from below the *SRT50n* point on

average have equal but opposite effects that are averaged out. In CI users, only a small bias (<0.85 dB) was present if optimal parameters are used. The bias depended on the relative initial SNR. An SNR more than 4 dB above the $SRT50n$ resulted in a relatively large positive bias. The explanation is that the slope of the intelligibility function well above $SRT50n$ becomes very shallow, making the adaptive procedure not very effective, as already explained for the npBK SA algorithm.

The within-staircase SD was dependent on the step size constant, the decrease rate of the step size, the number of trials, and the intelligibility function (s and p_{max}) of the group of listeners. As a consequence, the within-staircase SD cannot be used as a measure of the reliability of a single SRT50n measurement in combination with a fixed criterion (cf. Keidser et al., 2013). We analyzed if the SD and bias of the $SRT50n$ estimates was dependent on the within-staircase SD. In the stimulations, within-staircase SDs up to approximately twice the RMS within-staircase SD of the group were seen. For this range, no relationship was found for the averaged ndss SA with $b = 4$, neither in the CI group (Figure 4), nor in the NH group. This means that the within-staircase SD is not really suitable as a measure for the reliability of an individual staircase. Only if a single staircase has a very large within-staircase SD compared with the group value (as a rule of thumb: more than twice the RMS within-staircase SD of the group), one may decide to reject this measurement.

Influence of Maximum Intelligibility on Accuracy

A decrease of the maximum intelligibility in quiet p_{max} caused an increase in the SD of the $SRT50n$ estimates. This was as expected and was mainly caused by the decrease of the slope of the intelligibility function to p_{max} times the original slope at $p = \frac{1}{2} p_{max}$. At $p = 0.5$, the slope is reduced even more because at this point the slope is no longer at its maximum value. For a smaller part, the increase in the SD of the $SRT50n$ estimates was caused by a decreasing efficiency of the adaptive procedure for decreasing p_{max}. As can be seen from Figure 6, if p_{max} decreases, the difference between the SDs of the SA algorithms and the theoretical minimum SD increases. There was also some bias in the $SRT50n$ estimate, but this remained acceptable small (< 0.5 dB) if the initial SNR was not too far from the true $SRT50n$ value.

For CI users with $p_{max} \geq 0.7$, but < 1, it is advantageous to start at an SNR that is below the real $SRT50n$. Then, the trials are in the steepest part of the intelligibility function, which makes the SA algorithms converge better toward the target. As a result, both bias and SD were smaller (Figure 5). According to Figure 6, the minimum p_{max} required for reliable use of adaptive estimation of $SRT50n$ is $p_{max} = 0.7$ provided that at least 20 sentences are used.

The Simulation Model

The development and application of a realistic and detailed simulation model of speech recognition was an important part of this study. The usefulness of the model for single trials in adaptive procedures was verified by comparing the within-staircase SDs of the simulations with the within-staircase SDs of the participants in the studies that were used to determine the model parameters. They matched very well. Furthermore, simulation of sentence scoring was in good agreement with the data of Versfeld et al. (2000), and simulations of word scoring with the ndss SA for NH listeners agreed well with results of Dingemanse and Goedegebure (2019). These findings show that the model appears to be a valid tool for evaluation of adaptive speech-in-noise algorithms.

The good agreement between simulations and experimental data is based on the detailed and already validated model of Bronkhorst et al. (1993) that predicts the proportions correct of k out of l words correctly. In the model, the effect of contextual information is incorporated. Due to the contextual information, a listener has a higher chance to predict initial missed words correctly from the words that were already understood. Brand and Kollmeier (2002) also used Monte Carlo simulations to examine adaptive procedures for sentences-in-noise tests with word scoring. To account for the effect of the contextual information, they used the j factor of Boothroyd and Nittrouer (1988), a factor that quantifies the number of statistically independent words in a sentence. In their simulations, each trial consisted of j Bernoulli trials, and the proportion correct score for each trial was calculated by dividing the sum of the results of the Bernoulli trials by j. However, the resulting distribution of proportion correct scores is not in accordance with the distribution that is found in sentence recognition, having a relatively large proportion of 0 and 1 values (see Figure 1 and also Hu et al., 2015). Furthermore, only integer values of j can be used. In contrast, the multinomial distribution of proportions from the model of Bronkhorst et al. (1993) as shown in Figure 1 was in good agreement with experimentally found distributions for all percent correct values. Also noninteger values of j that were dependent of the proportion correct value were a result of this model (Dingemanse & Goedegebure, 2019).

We added small stochastic between-sentence variations in $SRT50n$ and slope that exist within speech materials and individual listeners. We also added between-subject variations in context parameters and slopes. Addition of these stochastic variations has made the model more realistic, but the effects of these variations were small. This is in accordance with the finding of Smits and Houtgast (2006), who also reported

that variations in *SRT50n* and slope had only a small effect in a digit-in-noise test.

In the simulation model, some lapsing was included, but the lapse rate was kept constant over time. In future use of simulation models, it is worth to consider more variation in this lapse rate to simulate variations in attention and/or fatigue. These variations should be based on experimental data on attention variations and fatigue effects. However, we expect that the effect of lapsing on the accuracy is limited. The effect of lapsing is comparable with a reduction of p_{max} (see Equation 6). Figure 6 shows that for a reduction of p_{max} from 1 to 0.9, the increase of the SD and bias of *SRT50n* was limited. So, for lapse rates smaller than 10%, the effect of lapsing on the *SRT50n* estimate is small.

Usefulness of Adaptive Speech-in-Noise Tests in CI Recipients

Although SA algorithms provide relatively accurate estimations of the *SRT50n* in CI users, the SD of the *SRT50n* estimate was still much larger in the CI group than in the NH group, depending on p_{max} and the slope of the intelligibility function. The decreased slope in CI users (even for $p_{max} = 1$) is due to difficulties in understanding the sentences in this open-set speech material with relatively good real-life similarity. In contrast, if a closed-set speech material is used, such as a matrix sentence test (Kollmeier et al., 2015), the difference in slope between CI and NH listeners is much smaller (Hey et al., 2014; Theelen-van den Hoek et al., 2014), and the *j* factor is higher: approximately 4 (Wagener et al., 1999). This may be of help to obtain a more reliable *SRT50n* value, but the ecological validity of the speech material is much less than the sentences used in this study.

The question is whether a larger SD of the SRT50n estimate in CI users is problematic. From the perspective of CI recipients, a perceived increase in speech intelligibility is more important than a change in *SRT50n*. If the slope of the intelligibility curve at 50% is shallow, a larger shift in SNR is needed to obtain a relevant increase in speech intelligibility. This allows a less accurate estimate of the SNR. A typical SD value for the SA procedures is 1.7 dB for 26 sentences of the speech material used in this study. An SNR difference of 1.7 dB corresponds to an intelligibility difference of 10%. In NH listeners, the SD of the SA methods is 0.6 dB, corresponding to an intelligibility difference of 9%. So, in terms of intelligibility, the accuracy of the speech-in-noise test in CI users is comparable with the accuracy in NH listeners.

Because of the relatively large SDs in the CI group, it is often not possible to compare two conditions or two algorithms within an individual. The test–retest SD is $\sqrt{2}$ times the SD of a single measurement. A significant difference at the .05 level requires a difference of at least

$1.96 \cdot \sqrt{2} \cdot$ SD. In our example, $1.96 \cdot \sqrt{2} \cdot 1.7 = 4.7$ dB. Therefore, only differences in conditions that result in large SRT differences can be reliably detected in individuals. If one wants to compare two conditions in a research setting, the relatively high SD can be compensated by the group size.

General Discussion and Limitations

In clinical practice, often the first sentence is presented repeatedly with increasing SNR until the sentence is recognized (Plomp & Mimpen, 1979). We also used this procedure in the simulations, but we used a relatively small step of 2 dB and restricted the number of repetitions to a maximum of 3. This restriction prevented for initial SNRs that are (much) greater than the *SRT50n* because these SNRs would have resulted in more variability in the *SRT50n* estimate (according to Figure 5). We recommend to make an educated guess of the *SRT50n* and to use this guessed *SRT50n* minus 2 to 4 dB as initial SNR. Such an educated guess may be based on norm data, preliminary data, a familiarization run, or on known relationships of the *SRT50n* with other clinically available speech recognition data, such as word scores (e.g., Gifford et al., 2008). Only if one has too little knowledge for an educated guess, it is better to use the procedure of repeating the initial trials at higher SNR (+2dB) with a maximum of three repetitions.

In this study, the target proportion correct was 0.5, regardless of the maximum speech intelligibility in quiet. Another option is to choose the target as half the maximum speech intelligibility in quiet. Then, the target is at the steepest part of the intelligibility function, and the function is more symmetrical around the target. This would lead to a smaller SD and bias for *SRT50n*. However, this option has three drawbacks: First, each participant is tested at his own target level, making it impossible to compare the *SRT50n* values among participants; second, the perceived difficulty of the test would become too high, which increases the risk that a participant gives up; third, the individual p_{max} must be measured beforehand.

This study has some limitations. First, the VU sentences were selected for equal intelligibility at sentence level in NH listeners and not at word level in CI listeners. We have taken this into account by making variations in SRT and slope per sentence in the simulation model, but this is only an approximation. Second, the search for the best adaptive procedure was only done with use of parameters for the context model and the intelligibility function which were derived from data obtained with the VU sentences. However, the context parameters of the VU sentences are expected to be comparable with other open-set sentence materials. For example, they are comparable with the context parameters of the Göttingen

sentence test reported by Bronkhorst et al. (2002). Only if a very different speech type is used, such as a matrix test (Kollmeier et al., 2015), it would be safer to repeat the simulations with a context model and an intelligibility function that are suitable to these materials.

To test if the results of this study are applicable to the matrix test, we did some simulations for matrix tests. The simulations were based on the context parameters of the Olsa test that were reported by Bronkhorst et al. (2002). For the intelligibility function, we used $p_{max} = 0.82$, and a slope of $13.5 \pm 4.6\%$/dB at $P_t = 0.5$, based on values of Hey et al. (2014). Simulations for a list length of 30 trials with the averaged ndss SA algorithm with $b = 4$ resulted in a test–retest SD of 0.75 dB, giving a 95% confidence interval of about 3 dB. This agrees well with the range of test–retest differences reported by Hey et al. in their Figure 3. This indicates that SA algorithms work well for the matrix test. In matrix tests, a maximum-likelihood estimation of $SRT50n$ is used. This estimation is computationally complex and may sometimes produce more than one maximum, especially if the number of sentences is small (Pedersen & Juhl, 2017). As an alternative, an SA algorithm could be used because SA algorithms are nonparametric and provide easy to calculate estimations of the $SRT50n$.

In this study, nonparametric SA algorithms were used to estimate the $SRT50n$. However, as discussed in the Introduction section, maximum-likelihood and Bayesian methods are also valuable options to estimate the $SRT50n$. Doire et al. (2017) reported on a robust Bayesian method and compared this method with the estimation methods of Brand and Kollmeier (2002) and Shen and Richards (2012). They reported simulation results for several psychometrical functions. One of these functions, having a slope of $0.075 \, dB^{-1}$ and a lapse rate of 0.1, is comparable with the simulations of the CI group in this study. In our study, the number of statistically independent trials for 26 sentences is 52 because the effective number of independent words in the VU sentences is 2 (Dingemanse & Goedegebure, 2019). Results of this study can therefore be compared with 52 trials in the Doire et al. study. For 52 trials, Doire et al. reported an SD of 2 dB and a bias of –1 dB for $SRT50n$ for all methods used. In this study, the values are better: $SD = 1.3 - 1.5$ dB, and the bias is around –0.5 to –0.3 dB (Figure 6 at $p_{max} = 0.9$). On the other hand, the method of Doire et al. may be more robust for initial SNRs that are relatively far from the true $SRT50n$. For future research, we recommend a comparison between the nonparametric SA methods, parametric maximum-likelihood-based methods, and Bayesian methods, all with the same listener simulation model as used in this study. Furthermore, more research is needed on how to extend the different methods to measure threshold, slope, and p_{max} concurrently.

Conclusions

In conclusion, this study showed that SA methods based on word scoring provide efficient estimations of the $SRT50n$ in sentence-in-noise measurements, both in CI recipients and in NH listeners, if used with optimized parameters that govern the step size sequence. Although intelligibility functions in CI users have less steep slopes and a lower maximum intelligibility score in quiet, SA algorithms are capable to estimate the $SRT50n$ efficiently. They have the advantage that knowledge of the maximum intelligibility score in silence and slope is not needed in the estimation of $SRT50n$.

The SA algorithms proposed in this study provided more efficient $SRT50n$ estimates than clinical used adaptive procedures. Therefore, they are recommended for clinical use. They may also lead to more statistical power of speech-in-noise tests if used in research or equivalently in a smaller number of participants that is needed to achieve sufficient statistical power.

The different SA algorithms used in this study provide equally accurate estimations of the $SRT50n$. This was found both for CI users and NH listeners. The averaged SA algorithm with a step size factor of 4 is recommended for clinical use because it is relatively easy, and it is applicable to a wide range of hearing-impaired listeners. In CI users, the most accurate estimate of $SRT50n$ is obtained if the initial SNR is chosen below the $SRT50n$, the step size is relatively small, and at least 20 sentences per condition are used. The within-staircase SD turned out not to be suitable as a measure for test reliability.

The SD of the $SRT50n$ estimate increases with decreasing maximum intelligibility in quiet. The score of words from sentences in quiet should be at least 70% correct for reliable use of adaptive estimation of $SRT50n$.

Author Contributions

G.D. designed the study and did the simulations. Both authors did the interpretation of the data. G.D. drafted the article, and A.G. revised the article. Both authors approved the final version of the article for submission.

ORCID iD

Gertjan Dingemanse (iD) https://orcid.org/0000-0001-8837-3474

References

Bather, J. A. (1989). *Stochastic approximation: A generalisation of the Robbins-Monro procedure*. Mathematical Sciences Institute, Cornell University.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America*, *84*(1), 101–114. https://doi.org/10.1121/1.396976

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, *111*(6), 2801–2810. https://doi.org/10.1121/1.1479152

Bronkhorst, A. W., Bosman, A. J., & Smoorenburg, G. F. (1993). A model for context effects in speech recognition. *The Journal of the Acoustical Society of America*, *93*(1), 499–509. https://doi.org/10.1121/1.406844

Bronkhorst, A. W., Brand, T., & Wagener, K. C. (2002). Evaluation of context effects in sentence recognition. *The Journal of the Acoustical Society of America*, *111*(6), 2874–2886. https://doi.org/10.1121/1.1458025

Chan, J. C., Freed, D. J., Vermiglio, A. J., & Soli, S. D. (2008). Evaluation of binaural functions in bilateral cochlear implant users. *International Journal of Audiology*, *47*(6), 296–310. https://doi.org/10.1080/14992020802075407

Dawson, P. W., Mauger, S. J., & Hersbach, A. A. (2011). Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients. *Ear and Hearing*, *32*(3), 382–390. https://doi.org/10.1097/AUD.0b013e318201c200

Dingemanse, J. G., & Goedegebure, A. (2015). Application of noise reduction algorithm ClearVoice in cochlear implant processing: Effects on noise tolerance and speech intelligibility in noise in relation to spectral resolution. *Ear and Hearing*, *36*(3), 357–367. https://doi.org/10.1097/AUD.0000000000000125

Dingemanse, J. G., & Goedegebure, A. (2018). Optimising the effect of noise reduction algorithm ClearVoice in cochlear implant users by increasing the maximum comfort levels. *International Journal of Audiology*, *57*(3), 230–235. https://doi.org/10.1080/14992027.2017.1390267

Dingemanse, J. G., & Goedegebure, A. (2019). The important role of contextual information in speech perception in cochlear implant users and its consequences in speech tests. *Trends in Hearing*, *23*, 2331216519838672. https://doi.org/10.1177/2331216519838672

Doire, C. S. J., Brookes, M., & Naylor, P. A. (2017). Robust and efficient Bayesian adaptive psychometric function estimation. *Journal of the Acoustical Society of America*, *141*(4), 2501. https://doi.org/10.1121/1.4979580

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology and Neuro-Otology*, *13*(3), 193–205. https://doi.org/10.1159/000113510

Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *Journal of the Acoustical Society of America*, *97*(6), 3749–3760. https://doi.org/10.1121/1.412390

Hagerman, B., & Kinnefors, C. (1995). Efficient adaptive methods for measuring speech reception threshold in quiet and in noise. *Scandinavian Audiology*, *24*(1), 71–77. https://doi.org/10.3109/01050399509042213

Hey, M., Hocke, T., Hedderich, J., & Muller-Deile, J. (2014). Investigation of a matrix sentence test in noise: Reproducibility and discrimination function in cochlear implant patients. *International Journal of Audiology*, *53*(12), 895–902. https://doi.org/10.3109/14992027.2014.938368

Hu, W., Swanson, B. A., & Heller, G. Z. (2015). A statistical method for the analysis of speech intelligibility tests. *PLoS One*, *10*(7), e0132409. https://doi.org/10.1371/journal.pone.0132409

Keidser, G., Dillon, H., Mejia, J., & Nguyen, C.V. (2013). An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function. *International Journal of Audiology*, *52*(11), 795–800. https://doi.org/10.3109/14992027.2013.817688

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, *29*(1), 41–59. https://doi.org/10.1214/aoms/1177706705

King-Smith, P. E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, *37*(12), 1595–1604. https://doi.org/10.1016/s0042-6989(96)00310-0

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*(Suppl 2), 3–16. https://doi.org/10.3109/14992027.2015.1020971

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. https://doi.org/10.1016/s0042-6989(98)00285-5

Kushner, H. J., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, *95*(2), 1085–1099. https://doi.org/10.1121/1.408469

Pedersen, E. R., & Juhl, P. M. (2017). Simulated critical differences for speech reception thresholds. *Journal of Speech, Language, and Hearing Research*, *60*(1), 238–250. https://doi.org/10.1044/2016_JSLHR-H-15-0445

Plomp, R., & Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology*, *18*(1), 43–52. https://doi.org/10.3109/00206097909072618

Polyak, B. T. (1990). New stochastic approximation type procedures. *Automation and Remote Control*, *7*(2), 98–107.

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on*

Control and Optimization, 30(4), 838–855. https://doi.org/10.1137/0330046

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3), 400–407. https://doi.org/10.1214/aoms/1177729586

Ruppert, D. (1988). *Efficient estimations from a slowly convergent Robbins-Monro process* (Technical Report No. 781). School of Operations Research and Industrial Engineering, Cornell University.

Schwabe, R. (1994). On Bather's stochastic approximation algorithm. *Kybernetika*, 30(3), 301–306.

Schwabe, R., & Walk, H. (1996). On a stochastic approximation procedure based on averaging. *Metrika*, 44(1), 165–180. https://doi.org/10.1007/bf02614063

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

Smits, C., & Houtgast, T. (2006). Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. *Journal of the Acoustical Society of America*, 120(3), 1608–1621. https://doi.org/10.1121/1.2221405

Soli, S. D., & Wong, L. L. (2008). Assessment of speech intelligibility in noise with the Hearing in Noise Test. *International Journal of Audiology*, 47(6), 356–361. https://doi.org/10.1080/14992020801895136

Terband, H., & Drullman, R. (2008). Study of an automated procedure for a Dutch sentence test for the measurement of the speech reception threshold in noise. *Journal of the Acoustical Society of America*, 124(5), 3225–3234. https://doi.org/10.1121/1.2990706

Theelen-van den Hoek, F. L., Houben, R., & Dreschler, W. A. (2014). Investigation into the applicability and optimization of the Dutch matrix sentence test for use with cochlear implant users. *International Journal of Audiology*, 53(11), 817–828. https://doi.org/10.3109/14992027.2014.922223

Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *Journal of the Acoustical Society of America*, 107(3), 1671–1684. https://doi.org/10.1121/1.428451

Vroegop, J. L., Dingemanse, J. G., Homans, N. C., & Goedegebure, A. (2017). Evaluation of a wireless remote microphone in bimodal cochlear implant recipients. *International Journal of Audiology*, 56(9), 643–649. https://doi.org/10.1080/14992027.2017.1308565

Wagener, K., Brand, T., & Kollmeier, B. (1999). Entwicklung und evaluation eines satztests für die deutsche sprache III: Evaluation des oldenburger satztests (Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test). *Zeitschrift Audiologie*, 38, 86–95.

Wong, L. L. N., & Keung, S. K. H. (2013). Adaptation of scoring methods for testing cochlear implant users using the Cantonese Hearing In Noise Test (CHINT). *Ear & Hearing*, 34(5), 630–636. https://doi.org/10.1097/AUD.0b013e31828e0fbb

Zhang, N., Liu, S., Xu, J., Liu, B., Qi, B., Yang, Y., . . . Han, D. (2010). Development and applications of alternative methods of segmentation for Mandarin Hearing in Noise Test in normal-hearing listeners and cochlear implant users. *Acta Otolaryngologica*, 130(7), 831–837. https://doi.org/10.3109/00016480903493758