





The one-up one-down adaptive (staircase) procedure in speech-in-noise testing: Standard error of measurement and fluctuations in the track^{a)}

Cas Smits,^{1,b)}  Joost M. Festen,² De Wet Swanepoel,^{3,c)}  David R. Moore,^{4,d)}  and Harvey Dillon^{5,e)} 

¹Amsterdam UMC location University of Amsterdam, Otolaryngology–Head and Neck Surgery, Ear and Hearing, Amsterdam Public Health Research Institute, Meibergdreef 9, Amsterdam, Netherlands

²Amsterdam UMC location Vrije Universiteit Amsterdam, Otolaryngology–Head and Neck Surgery, Ear and Hearing, Amsterdam Public Health Research Institute, De Boelelaan 1117, Amsterdam, Netherlands

³Department of Speech–Language Pathology and Audiology, University of Pretoria, Pretoria, Gauteng, South Africa

⁴Communication Sciences Research Center, Cincinnati Children’s Hospital Medical Center and University of Cincinnati, Cincinnati, Ohio 45229, USA

⁵Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, United Kingdom

ABSTRACT:

The one-up one-down adaptive (staircase or up-down) procedure is often used to estimate the speech recognition threshold (SRT) in speech-in-noise testing. This article provides a brief historical overview of the one-up one-down procedure in psychophysics, discussing the groundbreaking early work that is still relevant to clinical audiology and scientific research. Next, this article focuses on two aspects of the one-up one-down adaptive procedure: first, the standard error of measurement (SEM) and, second, the fluctuations in the track [i.e., the standard deviation of the signal-to-noise ratios of the stimuli within the track (SD_{track})]. Simulations of ideal and non-ideal listeners and experimental data are used to determine and evaluate different relationships between the parameters slope of the speech recognition function, SRT, SEM, and SD_{track} . Hearing loss and non-ideal behavior (inattentiveness, fatigue, and giving up when the task becomes too difficult) slightly increase the average value of SD_{track} . SD_{track} , however, poorly discriminates between reliable and unreliable SRT estimates. © 2022 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1121/10.0014898>

(Received 23 March 2022; revised 13 September 2022; accepted 20 September 2022; published online 25 October 2022)

[Editor: Matthew B. Winn]

Pages: 2357–2368

I. INTRODUCTION

The one-up one-down adaptive (staircase or up-down) procedure is probably the most common method to estimate the speech recognition threshold (SRT) in speech-in-noise testing. The adaptive procedure is used to provide an estimate of the true SRT. The random error in the estimate will be eliminated if the test is performed an infinite number of times or if the number of presentations in the test is infinite, but a systematic error (i.e., bias) can remain. The true SRT is defined as the signal-to-noise ratio (SNR) where a listener recognizes 50% of the speech items correctly. It corresponds to one point on the speech recognition function that relates the recognition probability to SNR. The speech recognition function is, in general, an S-shaped function that can be described by the SRT and a slope parameter and by two

additional parameters if the lower and upper asymptote deviate from 0 and 1, respectively. The one-up one-down adaptive procedure is implemented in many standard speech-in-noise tests (e.g., Plomp and Mimpen, 1979; Nilsson *et al.*, 1994; Cameron and Dillon, 2007; Smits *et al.*, 2013).

As for all outcome measures, an essential requirement of the SRT is that it be valid and reliable (Mokkink *et al.*, 2010). The standard error of measurement (SEM, or measurement error) is an agreement parameter and quantifies how close the results for repeated measurements in one listener are. The SEM is needed to determine whether two SRTs are significantly different or whether an SRT is significantly different from a certain value. When using the one-up one-down adaptive procedure to estimate the SRT for a listener, the SEM of this estimate depends on several factors. First, a steeper slope of the speech recognition function yields a more precise SRT estimate, thus, a smaller SEM, than a shallow slope (Theunissen *et al.*, 2009). Ensuring a steep speech recognition function is one of the most important aspects of the development of speech-in-noise tests. Second, several parameters of the measurement procedure, for example, the step size, number of presentations, and starting level, affect the SEM (García-Pérez, 1998; Smits and Houtgast, 2006).

^{a)}This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.

^{b)}Electronic mail: c.smits@amsterdamumc.nl

^{c)}Also at: Ear Science Institute Australia, Subiaco, Australia.

^{d)}Also at: Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, United Kingdom.

^{e)}Also at: Department of Linguistics, Macquarie University, Sydney, Australia.

Third, listener factors play a role. The SEM will generally be larger in listeners with hearing loss than in normal-hearing listeners (Smits and Festen, 2011). In addition, inattentiveness or other non-ideal behavior can negatively affect the SEM. The present paper focuses on the first and third factor, where we will pay specific attention to the information contained in the fluctuations in the adaptive track and how this relates to the SEM.

The SEM can be estimated from the standard deviation (SD) of the values of repeated measurements of one listener. When two measurements for each of n listeners are available (e.g., SRT_{test} and SRT_{retest}), the average SEM for this group of listeners can be determined from the distribution of the differences ($diff = SRT_{test} - SRT_{retest}$) between test and retest values (Plomp and Mimpen, 1979; Smits and Houtgast, 2005; de Vet et al., 2006),

$$SEM = \frac{SD(diff)}{\sqrt{2}}. \tag{1}$$

Note that a systematic error is not included in this calculation. Thus, for example, a systematic learning effect (i.e., equal learning effect for each listener) has no effect on the SEM determined by Eq. (1). When including the systematic difference in the agreement parameter SEM, the following equation can be used:

$$SEM = \sqrt{\sum_{i=1}^n \frac{diff_i^2}{2 \cdot n}}. \tag{2}$$

A similar and related difference exists between the intraclass correlation coefficients (ICCs) $ICC_{agreement}$ and $ICC_{consistency}$ (de Vet et al., 2006). Smits et al. (2004) suggested a procedure to estimate the SEM from single SRT measurements when no test-retest data are available. Their procedure can be considered as a modified split-half method. The SEM can be used to generate confidence intervals around measured SRTs. In general, the SEM is determined from test and retest measurements in a (large) group of listeners. It means that this SEM is accurate as an average value for a population, but it may fail to reflect the variability of its value. For example, it has been demonstrated that the SEM is not constant for speech-in-noise measurements but increases with SRT (Smits and Festen, 2011).

The use of a specific and reliable SEM, based on an individual measurement, would be preferable to account for variability in the SEM due to hearing loss or unreliable responses from the listener. This information is very desirable in clinical settings where repeating measurements is less common than in scientific studies; but also in scientific studies, it would be helpful to have objective criteria to exclude unreliable SRT estimates. To the best of our knowledge, Bode and Carhart (1973) were the first to suggest using the standard error (SE) of the variation in presentation levels as an estimate of the precision of a single test. Unfortunately, they did not provide evidence for their suggestion, which has been widely used by clinicians and

researchers without the necessary validation (e.g., Koole et al., 2016; Jacobi et al., 2017; Sheikh Rashid et al., 2017; Denys et al., 2018). Its popularity is probably based on the intuitive belief that SRTs based on tracks with small fluctuations are more reliable than tracks with larger fluctuations. Here, fluctuations refer to the range of SNRs of the presentations. The fluctuations in a track can be quantified by the SD of the SNRs of the stimuli within the track (called SD_{track} in the current paper), and the assumption follows that SD_{track} is related to the reliability of the SRT estimate calculated from that specific track. It has typically been assumed that SRT estimates from highly fluctuating tracks are more likely to deviate from the true SRT or that the SEMs of these SRT estimates are larger than those from tracks with small fluctuation. The use of SD_{track} as a reliability measure is loosely based on the notion that the SEM is related to the SD of independent repeated measurements. However, SD_{track} is essentially different because the SNRs within a track are highly dependent; thus, the use of SD_{track} as a reliability measure should be examined.

In this paper, a description and definitions of the one-up one-down procedure will be presented, followed by a brief historical overview. Without striving for completeness, our goal is to trace the seminal early work that still has relevance to the clinical and research endeavors in speech-in-noise testing. In the remainder of this paper, we will focus on the two abovementioned aspects of the one-up one-down adaptive procedure: (1) the relationship between slope of the speech recognition function and SEM and (2) the interpretation and relevance of the SD of the SNRs of the stimuli within the track, SD_{track} , using simulations and experimental data from listeners.

A. The one-up one-down adaptive procedure

Essentially, in the adaptive procedure, the SNR of each presentation is based on the correctness of the previous response. That is, if the presentation is recognized correctly, the next presentation is presented at a lower SNR, and if the response is incorrect, the next presentation is presented at a higher SNR. The levels of the presentations follow a track. Figure 1 provides an example of a track and overview of the definitions used to describe the procedure. Three important parts of the track can be identified: the initial presentation,

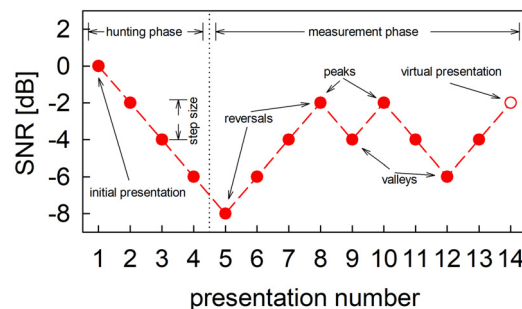


FIG. 1. (Color online) Example of an adaptive track and the definitions used to describe the one-up one-down adaptive procedure.

the hunting phase, and the measurement phase. First, the starting level of the initial presentation must be selected. A level above the expected SRT is commonly chosen. Then we have the next part, aptly referred to as the hunting phase in the manual of the speech-in-noise test of the UK Biobank (2012) study. The main aim is to get close to the true SRT in this part. It consists of a fixed number of presentations, or, for example, it ends after the first incorrect response (i.e., the first reversal). A secondary aim can be to familiarize the listener with the task. The last part is the actual measurement phase. The SRT estimate is based on presentation levels in this part. The difference in SNR between two consecutive presentations is the step size, d , of the procedure. The step size is often constant during the task, and the SRT is determined by averaging the SNRs of the presented stimuli, or the SNRs of the reversals, while omitting the initial presentation and the presentations in the hunting phase. Sometimes different step sizes are used in this phase. Advantages of the method compared to, for example, maximum-likelihood methods are the insensitivity to lapses (unforced errors due to inattention or accidentally entering the wrong response) (Smits and Houtgast, 2006), the lack of necessity to know the exact shape of the speech recognition function, and the small effect of the choice of step size on the SRT estimate.

B. Brief historical overview on the staircase procedure in speech-in-noise testing

The up-down method does not originate from research of speech-in-noise testing or psychophysics. Dixon and Mood (1948) introduced the method for obtaining sensitivity data. They discuss—as an example—the method for an experiment in which the sensitivity of explosives to shocks is tested. The first explosive is tested by dropping a weight from a certain height. If it explodes, the next weight is dropped from a lower height; otherwise, it will be dropped from a higher height. Dixon and Mood (1948) provided analytical, approximate maximum likelihood, estimates for the mean and SD of the threshold based on test data. Analysis required that the threshold be normally distributed, the sample size be large, and the interval between testing levels (the step size) be approximately equal to the SD of the threshold. The method became increasingly popular in different areas like the estimation of the median lethal dose (LD_{50}), fatigue testing, and psychophysics. Brownlee *et al.* (1953) proposed a much easier method to determine the mean. They suggested using the arithmetic mean of the n presentations used. They also suggested the inclusion of the $(n + 1)$ st presentation, which is not presented (a virtual presentation, the level of which is based on the correctness of the preceding presentation) in the calculation. The procedure gained popularity in audiological research, starting with the publication from Levitt and Rabiner (1967) and continuing with Levitt's classic paper on the transformed up-down method (Levitt, 1971). The method to estimate the mean, proposed by Levitt (1971) and Wetherill and Levitt (1965), is different from the estimators of Dixon and Mood (1948) and Brownlee *et al.*

(1953), as they suggest averaging the peaks and valleys (reversals) of the track. This procedure is slightly more precise than that of Brownlee *et al.* (1953), in which all presentation levels are averaged (Wetherill *et al.*, 1966). It was proven that the up-down procedure is highly efficient, but only when the initial presentation is within a few steps of the true mean. If the first stimulus is too high or too low, the responses to the stimuli will be all correct or incorrect until the track reaches the region of the true mean. Including these responses in the estimator will give a poor estimate of the true mean. Brownlee *et al.* (1953) suggested omitting the initial run of responses of the same sign in the calculation of the estimate. That is, the measurement phase starts with the first reversal.

Whereas in the earlier research (e.g., Dixon and Mood, 1948), the up-down method was presented as a way to calculate the mean and SD of a variate analytically, its use in the speech-in-noise literature is mainly limited to the determination of the mean (i.e., the SRT). The SD of the variate in the context of speech-in-noise testing is, however, an important parameter because it describes the slope of the speech recognition function.

To determine and evaluate various relationships between parameters (track length, slope of the speech recognition function, SRT, SEM and SD_{track}), both simulations and experimental data were used in the present study.

II. METHODS

We simulated ideal listeners and captured inattentiveness and other human factors in simulated non-ideal listeners. Additionally, we analyzed experimental data to verify some of the results from simulations. In the simulations, a brute force calculation model was used for relatively short track lengths, whereas for longer track lengths, Monte Carlo simulations were used. Three datasets from previously published studies (Smits and Houtgast, 2005; Smits *et al.*, 2013; Smits *et al.*, 2016) were reanalyzed.

A. Simulations

The brute force calculation model of Smits and Houtgast (2006) was used to perform exact calculations for speech-in-noise tests with 13 presentations per track. We decided to use this number of presentations because the simple adaptive up-down method with a step size of 2 dB and using 13 presentations has been very common in speech-in-noise tests since it was first proposed by Plomp and Mimpen (1979). Including the virtual 14th presentation and omitting the first four presentations (the initial presentation and the presentations in the hunting phase) yields a total of ten presentations in the measurement phase ($n = 10$). Brute force means that systematically all possible tracks are investigated, and the SRT, SD_{track} , and associated probability for each track are calculated. Then the weighted average SRT, SEM, and SD_{track} can be calculated. The advantage of this method is that the outcomes are error-free; the disadvantage is the exponential increase in number of tracks with track

length, which makes the method less suitable for longer track length due to the computational impact of the calculations.

For longer tracks (i.e., more than 13 presentations per track) Monte Carlo simulations were used with 10 000 runs for each set of parameters. Tests with 23 presentations per track were simulated except when exploring the effect of track length. The SRT estimate was calculated by averaging the SNRs of presentation 5 to the virtual 24th presentation ($n = 20$).

The speech recognition function, $\Phi(\text{SNR})$, can be any arbitrary S-shaped function, but traditionally a cumulative normal distribution is used. With a lower asymptote (i.e., the guess rate), γ , and upper asymptote $1 - \lambda$, defined by the lapse rate (or miss rate), λ , this function (Green, 1995) is described by

$$\Phi(\text{SNR}) = \gamma + (1 - \gamma - \lambda) \cdot \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\text{SNR}} e^{-(\zeta - \text{SRT})^2 / 2\sigma^2} d\zeta. \quad (3)$$

When $\gamma = \lambda = 0$, the slope of the function is maximal at $\text{SNR} = \text{SRT}$, and this slope value, S_{50} , expressed in %/dB, can be derived from σ by

$$S_{50} = \frac{100}{\sigma\sqrt{2\pi}}. \quad (4)$$

See, for example, Smits and Houtgast (2006) for further details and the effect of guess rate and lapse rate on slope values.

We simulated ideal normal-hearing listeners, ideal listeners with hearing loss, and non-ideal listeners. Although it seems unrealistic to assume that human listeners show very specific and isolated non-ideal behavior, we decided to use such an approach in our simulations. This allows us first to separate the effects of different forms of non-ideal behavior. Second, we can ascertain if, given a certain SRT and value of SD_{track} , non-ideal behavior could potentially be identified. Note that the downside of this approach is that, for example, it is likely more realistic that fatigued listeners will also experience lapses in addition to the steady shift in SRT. Figure 2 describes how different groups of listeners were modeled. The true SRT of the normal-hearing listeners was -10 dB SNR. The parameters for the other groups were chosen to yield average SRTs of -7 dB SNR for each group. We modeled the speech recognition function with a cumulative normal distribution [Eq. (3)]. To avoid relevant effects of the starting SNR on the SRT estimate (Smits and Houtgast, 2006), a starting SNR equal to the SRT of normal-hearing listeners was selected. This value is near to the SRT of all listeners given the four presentations in the hunting phase. Speech recognition functions for hearing-impaired listeners are shallower than for normal-hearing listeners. This can be understood when one realizes that, for hearing-impaired listeners, part of the signal is inaudible and/or contains less useful speech information because of the suprathreshold deficits these listeners may have. Then a

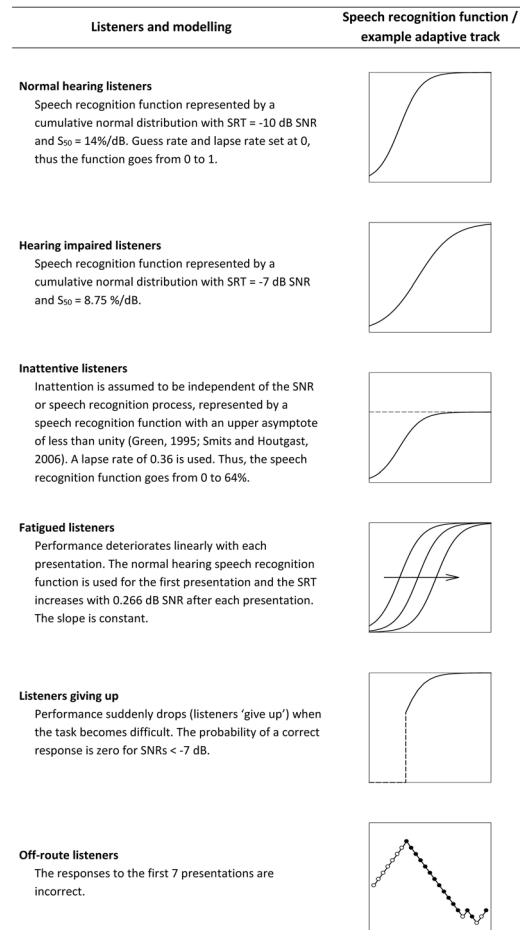


FIG. 2. Summary of how the different groups of listeners are modeled. Simulated listeners are ideal normal-hearing listeners, ideal listeners with hearing loss, and non-ideal listeners.

1-dB change in SNR will result in a smaller change in available speech information than for normal-hearing listeners. Therefore, the slope of the hearing-impaired speech recognition function can be predicted from the slope of the normal-hearing speech recognition function (Smits and Festen, 2011). The off-route listeners represent listeners who, for example, do not get close to the SRT before the presentations are given in the measurement phase (i.e., during the hunting phase) or do not respond correctly at all during a successive series of presentations (high lapse rate for some period). If the starting SNR is too far from the SRT or the listener unintentionally responds incorrectly to a few presentations, the SNRs of the following presentations are relatively favorable, and a sequence of correct responses may follow. This could result in a bias in the SRT estimate (Smits and Houtgast, 2006).

B. Experimental data

The dataset from the Dutch National Hearing test contains 39 968 SRTs. These are data from callers to the Dutch National Hearing Test. This test measured the SRT by telephone using digit triplets as speech material. Series of 23

triplets were used per SRT estimate ($n = 20$). Details have been reported (Smits and Houtgast, 2005, 2006).

Smits *et al.* (2013) describe the development of the Dutch Digits-In-Noise (DIN) test. In one of the experiments, they assessed a possible learning effect in the DIN test by measuring 25 SRTs for each of ten normal-hearing participants. When omitting the first SRT, no further learning effects were observed in 24 subsequent SRT measurements.

III. RESULTS

A. The SEM

1. Relationship between slope of the speech recognition function and SEM

To the best of our knowledge, there is no mathematically derived equation that describes the relationship between SEM, the parameters of the up-down procedure, and the speech recognition function. The exception is an up-down procedure where the speech recognition function can be described by a cumulative normal distribution with slope, S_{50} (i.e., the slope at 50% correct). Then for large values of the number of presentations, n , the dependence of SEM on S_{50} equals (Brownlee *et al.*, 1953; Wetherill, 1963)

$$SEM = \frac{G}{S_{50}\sqrt{n\pi}}, \tag{5}$$

with $G = 1$ when the step size, d , is related to S_{50} by

$$d = \frac{1}{S_{50}\sqrt{2\pi}}. \tag{6}$$

This step size was suggested by Dixon and Mood (1948) for their procedure.

Thus, the SEM is inversely proportional to the slope of the speech recognition function and the square root of the number of presentations. It is interesting to compare this value to the theoretically minimum SEM (Wetherill, 1963; Taylor, 1971),

$$SEM_{min} = \frac{1}{2 \cdot S_{50}\sqrt{n}}. \tag{7}$$

Thus, $SEM \approx 1.13 \times SEM_{min}$ for this step size, which illustrates how precise the one-up one-down adaptive procedure is.

In practice, often the chosen step size does not follow Eq. (6) but is smaller. The recommended step size for the one-up one-down adaptive procedure is related to the (often unknown) slope of the speech recognition function and is between 0.5 and 1 times $1/(S_{50}\sqrt{2\pi})$ (Wetherill, 1963). When the step size does not exactly fulfill Eq. (6), then G in Eq. (5) is not exactly equal to 1 (Dixon and Mood, 1948; Wetherill, 1963). We ran Monte Carlo simulations with varying step sizes and numbers of presentations to determine the value of G . Figure 3 shows G as a function of $d \times S_{50}$ for three different numbers of presentations per test. G clearly depends on $d \times S_{50}$, but most values are between

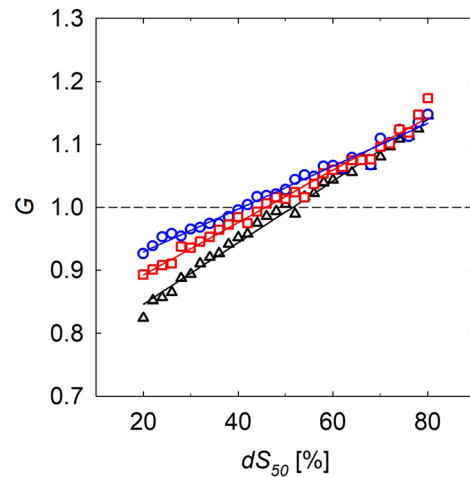


FIG. 3. (Color online) The value of G as a function of $d \times S_{50}$ for the one-up one-down adaptive procedures with $n = 100$ (blue circles), $n = 20$ (red squares), and $n = 10$ (black triangles). G represents the correction factor needed to calculate SEM from Eq. (5) when the step size and slope of the speech recognition function do not fulfill Eq. (6).

0.9 and 1.1. Thus, it can be concluded that the SEM is inversely proportional to the square root of the number of stimuli in the track and the slope of the speech recognition function. The exact value can be calculated with Eq. (5) when knowing G , the value of which depends on the product of step size and slope of the speech recognition function.

2. Use of a modified split-half method to estimate the SEM

Smits *et al.* (2004) suggested a procedure to estimate the SEM from a single SRT measurement. Their procedure can be considered as a modified split-half method. The presentations from the measurement phase are split in two. Then the first and second half are used as independent SRT estimates. However, these SRT estimates are not fully independent, and the calculated SEM is smaller than the true SEM. To make this clear: If the SRT estimate based on the first half is higher than the true SRT, then the probability that the SRT estimate based on the second half is also higher than the true SRT is over 50%. The main reason is that it is likely that the starting level of the second half will be above the true SRT, resulting in a bias in the SRT estimate. Thus, it must be concluded that the suggested procedure systematically underestimates the true SEM, and test-retest measurements are needed to determine the exact value of the SEM. Note that this is a “small sample effect,” which diminishes with the number of presentations. We ran Monte Carlo simulations, and the results show that for a set of realistic parameters (24 presentations of which 4 are in the hunting phase, $S_{50} = 14\%/dB$ and step size = 2 dB), the true SEM is approximately 12% higher than the SEM calculated from the modified split-half method (0.84 vs 0.75 dB). The exact factor depends on the parameters used; thus, it is not possible to estimate SEM in an unbiased manner using any single correction factor.

3. Use of SD_{track} to estimate the SEM

When a measurement is repeated n times in one listener, and those measurements are independent of each other, the SEM can be estimated as SD/\sqrt{n} . Cameron and Dillon (2007) noted that the individual presentation levels of an adaptive track are not independent but proposed (and used) SD_{track} to estimate the SEM with a multiplier of 2.0 to allow for the lack of independence. Thus, $SEM = 2 \cdot SD_{\text{track}}/\sqrt{n}$. The choice of multiplier was based on Monte Carlo simulations performed with a slope of the speech recognition function of 15%/dB and a step size of 2 dB. To examine the accuracy of this method of estimating SEM in the standard one-up, one-down procedure, we carried out Monte Carlo simulations with the slope ranging from 5%/dB to 25%/dB, $n = 20$, and a step size of 2 dB. Figure 4 shows that, when the slope is 15%/dB, the average value of the estimated SEM (0.84 dB) is, not surprisingly, very close to the true value of 0.85 dB, obtained by calculating the SD of the SRTs of each run. Note that for this slope value, $2 \cdot SD_{\text{track}}$ equals $1/(S_{50} \cdot \sqrt{\pi})$ from Eq. (5); both values are approximately 3.76. As would be expected, lower true slopes (i.e., S_{50}) result, on average, in a larger estimated SEM. However, for such slopes, which are precisely those that result in poorer measurement accuracy, the procedure systematically underestimates the SEM. The reason that SEM values calculated from SD_{track} using the procedure from Cameron and Dillon (2007) deviate from the true SEM is that they propose a constant factor of 2 in their calculation. Monte Carlo simulations showed that this factor should vary with slope (Keidser et al., 2013), but they choose a constant factor because the underlying slope is unknown. Further, even when the slope is around the expected value of 15%/dB, the SD of the estimated SEM, represented by the error bars in Fig. 4, is relatively large. The dashed red line in Fig. 4 represents Eq. (5) with $G = 1$. The true SEM deviates from this

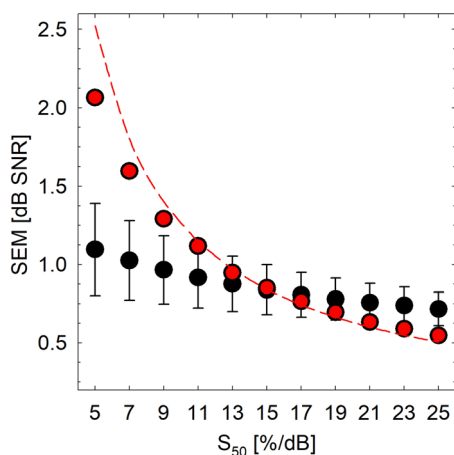


FIG. 4. (Color online) True SEM and estimated SEM vs the slope of the underlying speech recognition function. Black circles show the mean, and error bars the SD, of the SEM calculated using $2 \cdot SD_{\text{track}}/\sqrt{n}$, where n is the number of presentations. The red circles show the true SEM, calculated as the SD of the SRTs obtained from each run. The dashed red line represents the calculated SEM using Eq. (5) with $G = 1$.

line because the step size is 2 dB for all slope values. Then G should vary with S_{50} to compensate for this (see Fig. 3).

B. The (near) ideal listener: Fluctuations in the track, SD_{track}

1. Relationship between SD_{track} and SRT estimates

Many people feel that a perfect staircase (i.e., alternating correct and incorrect response) gives a better estimate of the true SRT than a strongly fluctuating track, even in an ideal listener. They might consider a retest for strongly fluctuating tracks and accept results from less fluctuating tracks. In this section, we evaluate whether high values of SD_{track} are related to either poorer SRTs on average or less accurate estimates of SRT. The evaluation is in the context of all listeners having identical (for the simulation) or very similar (for experimental data) speech recognition function slopes.

a. Simulations. The brute force calculation model was used to calculate the distribution of SD_{track} values for repeated measurements in an ideal listener. Typical parameters for the speech recognition function (cumulative normal distribution with a slope of 14%/dB), step size of 2 dB, and a true SRT of -10 dB SNR were used in the simulations. These parameters were fixed, thus, simulating an ideal listener or homogeneous group of listeners. The SRT estimates were ranked according to the value of SD_{track} . Then ten percentile groups were created such that the sum of the probabilities of all tracks in each group equaled 0.1. Next, the weighted average SRT, the SD of the SRTs (i.e., the SEM), and the weighted average SD_{track} values were calculated for each group. The results are shown in Fig. 5(A). They clearly demonstrate that there is no relationship between SD_{track} and the average SRT or the SD of the SRTs. The average SRT and variance in SRT for the “perfect” tracks (i.e., alternating between correct and incorrect responses; the low number groups) are equal to these values for the highly fluctuating tracks (the high number groups). Thus, the tracks with a larger SD_{track} do not provide less accurate SRT estimates than the tracks with a smaller SD_{track} , and, therefore, it can be concluded that SD_{track} cannot be used as a measure of inaccurate SRT estimates, at least when all estimates come from tracks that were formed using the same underlying speech recognition function.

b. Experimental data. Smits et al. (2013) reported no significant differences between 24 subsequent SRT measurements in ten listeners. We calculated SD_{track} for each SRT estimate and ranked the 24 SRTs for each listener according to SD_{track} . Figure 5(B) shows the mean and SD of the SRTs across listeners for each rank number. Also shown are the mean values of SD_{track} for each rank number. The results show a clear increase in SD_{track} , but the mean SRTs are all around the average value of -10.0 dB SNR. Linear regression analysis was used to assess the relationship between rank number and mean SRT in greater detail. The slope of the regression line did not significantly differ from

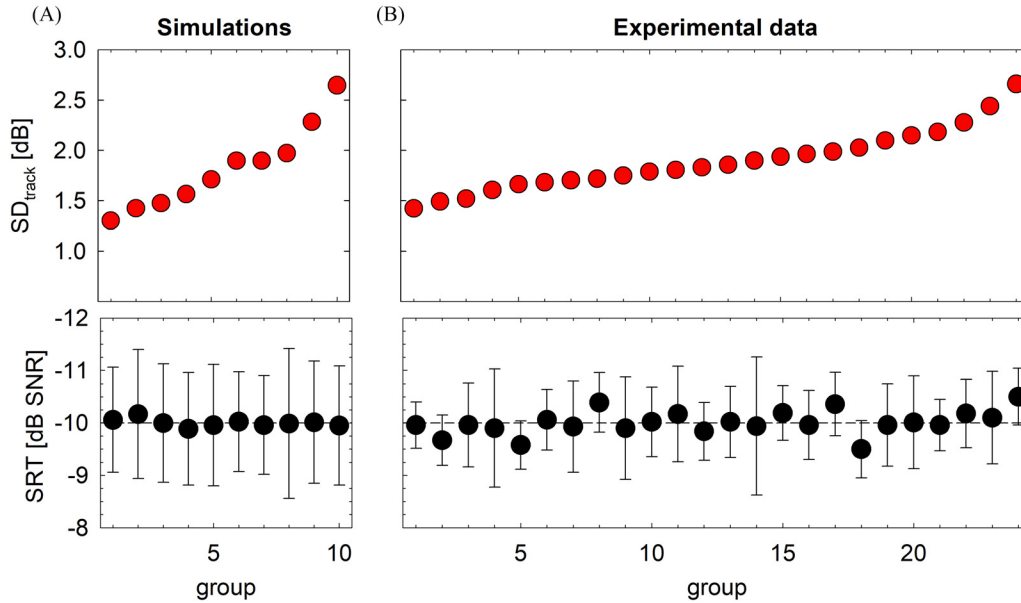


FIG. 5. (Color online) (A) Results from brute force calculations. Shown are mean and SD of SRT estimates (lower panel) and mean SD_{track} (upper panel) for groups based on SD_{track} values. (B) Results from 24 repeated measurements in ten normal-hearing listeners. For each listener, SRTs were ranked according to the SD_{track} . In black, the mean and SD of SRT estimates for each of these groups are ranked from lowest SD_{track} to highest SD_{track} . In red is shown the mean value of SD_{track} for each group.

0 ($\beta = -0.365$, $p = 0.08$). Analyses of a different dataset (Smits *et al.*, 2016) confirm this finding (see supplementary material¹).

2. Relationship between SD_{track} and slope of the speech recognition function

The probability that SNRs of the presentations within a track deviate from the true SRT is higher for shallow speech recognition functions than for steep speech recognition functions. This implies that the average SD_{track} will be higher for shallower speech recognition functions. Consequently, misinterpretation of highly fluctuating tracks can easily happen.

a. Simulations. The brute force calculation model was used to determine the relationship between SD_{track} and slope, S_{50} , of the speech recognition function for the adaptive procedure with 13 presentations ($n = 10$) and a step size of 2 dB. The relationship was also determined for tracks with 23 presentations ($n = 20$) with Monte Carlo simulations. Both for the brute force method and the Monte Carlo simulations, the slope value, S_{50} , was systematically changed between 1 and 25%/dB in steps of 1%/dB. The results are shown in Fig. 6(A), and they demonstrate the non-linear relationship between SD_{track} and S_{50} .

b. Experimental data. Smits and Houtgast (2006) explored the effect of SRT on the slope of the speech recognition function for a set of almost 40 000 SRTs. Groups were created for a grid of rounded SRTs and age (6 SRT groups \times 8 age groups). Average speech recognition functions were constructed for each group based on approximately 15 000 presentations per group. It was concluded

that the slope of the speech recognition function decreases with increasing SRT, independent of age. We calculated the average SD_{track} for each group. Figure 6(B) shows the slope of the speech recognition against SD_{track} . Each color represents groups with equal SRTs. It demonstrates that listeners with poorer SRTs have shallower speech recognition functions and higher values for SD_{track} . This result is as expected because the slope of the speech recognition function is related to the SRT, with shallower slopes for poorer SRTs (Smits and Festen, 2011). Note that, because the data were corrected for interindividual differences in SRT, the calculated slope values are greater than the true slope values (Smits and Houtgast, 2006). The experimental data confirm

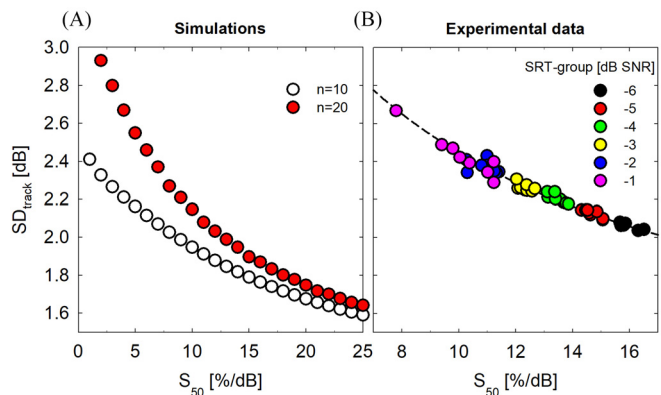


FIG. 6. (Color online) Relationship between the slope of the speech recognition function and SD_{track} . (A) Results based on simulations for 13 and 23 presentations and a step size of 2 dB. (B) Results from experimental data. Each circle represents data from a group of listeners with similar age and rounded SRT.

the conclusion from the simulations that, on average, SD_{track} is a measure of the slope of the speech recognition.

C. Non-ideal listeners: The effect of unstable or unreliable responses

The simulations in Sec. III B for (near) ideal listeners showed that fluctuations in the adaptive track, represented by SD_{track} , are of no value to determine whether an SRT estimate is likely to be near the true SRT, when all SD_{track} values originate from the same speech recognition function. The SRT estimates from highly fluctuating tracks were as reliable as SRT estimates from little fluctuating tracks. The experimental data confirmed these findings for several studies with adult participants. However, it cannot be ruled out that the results are different for special populations like young children, where, for example, procedural learning, attention, task difficulty, and the presence or absence of an experimenter may play a larger role. We therefore identified several “groups of listeners” that represent a typical non-ideal behavior, modeled this behavior, and used Monte Carlo simulations to explore the effect on the SRT and SD_{track} . The results were compared to Monte Carlo simulations for ideal listeners with normal hearing or hearing loss. The model parameters for each group were chosen to ensure that the mean SRT for the group was 3 dB poorer than that of the normal-hearing group (see Sec. II A and Fig. 2 for details).

1. Simulations

The panels in Fig. 7(A) show the distributions of SRTs and SD_{track} for the different groups. For clarity, only 200 datapoints are shown in each panel, and a small jitter (between ± 0.1 dB) was applied to avoid too much overlap of datapoints. The gray areas represent the mean $\pm 2SD$ for the normal-hearing listeners’ SRTs, thus, covering 95% of these listeners. The mean SRTs for the hearing-impaired listeners and the other groups of listeners are all approximately -7 dB SNR (ranging from -7.1 dB SNR for the hearing-impaired listeners to -6.8 dB SNR for the off-route listeners). The vertical dashed line represents a cutoff value for SD_{track} of 3.1 dB. This is a somewhat arbitrarily selected value, but it means that approximately 5% of the SRTs of the hearing-impaired listeners are incorrectly classified as unreliable. Note that the mean of these “unreliable” SRTs equals -7.1 dB SNR as well. The panels representing results from normal-hearing listeners with unstable or unreliable responses (upper four panels) show that the effect on SD_{track} is relatively small except for the off-route listeners (top panel). Most of these listeners have elevated SRTs but SD_{track} values below the cutoff value. This is quantified in Fig. 7(B), which shows the percentage of SRT estimates that would pass a certain cutoff value for SD_{track} (i.e., the pass rate). Ideally, the pass rate would be 100% for the ideal normal-hearing and hearing-impaired listeners (lower two panels) and 0% for the non-ideal listeners (upper four panels). The panels in Fig. 7(B) clearly show that no reasonable compromise is available between accepting reliable SRT

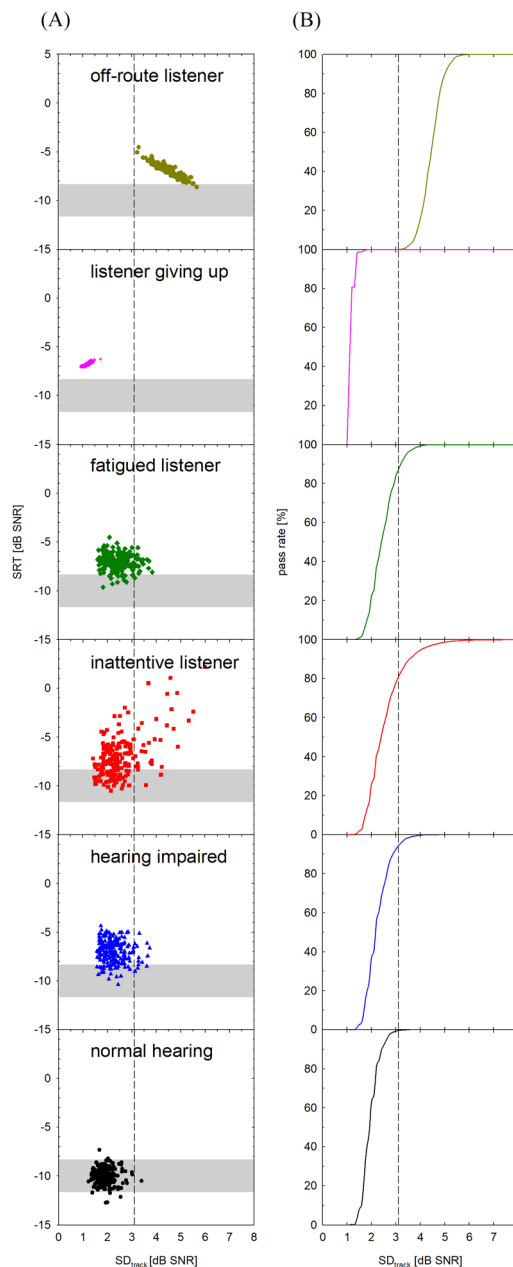


FIG. 7. (Color online) Results from different simulated listeners. The top four rows represent non-ideal listeners, the fifth row represents ideal hearing-impaired listeners, and the bottom row represents ideal normal-hearing listeners. The parameters were chosen such that the average SRT for normal-hearing listeners equals -10 dB SNR, and that for the other groups of listeners is -7 dB SNR. (A) SRT against SD_{track} . Only 200 datapoints from 10 000 simulations are shown per panel. A small amount of jitter (between ± 0.1 dB) was applied to the datapoints to avoid too much overlap. The gray areas represent the mean $\pm 2SD$ for the normal-hearing listeners. (B) Percentage of SRT estimates that would pass a certain cutoff value for SD_{track} (i.e., the pass rate). Ideally, the pass rate would be 100% for the ideal normal-hearing and hearing-impaired listeners (lower two panels) and 0% for the non-ideal listeners (upper four panels).

estimates and rejecting unreliable SRT estimates. The only exceptions are the off-route listeners, who can be identified by using a cutoff SD_{track} of approximately 3.5 dB for this track length, slope of the speech recognition function, and step size.

IV. DISCUSSION

The simple one-up one-down adaptive procedure has been popular in psychophysical research for many decades. It is probably the most widely used procedure in both tone and speech-in-noise testing. The elegance of the procedure lies in the simplicity of test administration and calculation of the threshold; no assumptions have to be made about the underlying speech recognition function, except that it increases monotonically.

A. Aspects of the SEM

The SEM is inversely proportional to both the slope of the speech recognition function and the square root of the number of presentations as shown in Eq. (5). It emphasizes the importance of a steep speech recognition function for the stimuli used in a speech-in-noise test. A prerequisite for this is that the recognition probabilities of the individual stimuli are equal, that is, the speech material should be homogeneous. According to the probabilistic model of Kollmeier (see, e.g., Zokoll *et al.*, 2012), the slope of the average speech recognition function of all stimuli in the test can be derived from the variance between recognition probabilities of the individual stimuli. It emphasizes the importance of equalizing recognition probabilities of stimuli in a speech-in-noise test (Dillon, 1983).

The current paper shows that the modified split-half method as suggested by Smits *et al.* (2004) and used in several studies (e.g., Smits and Houtgast, 2005; Denys *et al.*, 2018) cannot be used to accurately determine the SEM because it underestimates the true SEM. A correction factor could improve the accuracy of the modified split-half method, but a within-subjects (repeated measures) design is a preferable, unbiased method to determine the SEM of a test. It then remains to be decided whether to include a possible learning effect in the calculation of the SEM [Eq. (2)] or to exclude the systematic differences and use Eq. (1). Alternatively, the SEM can be estimated from Eq. (5) when the slope of the speech recognition function, for the population in question, is known. Whatever the choice, it is important to indicate the rationale behind it and to clarify how the SEM was calculated when publishing.

B. Fluctuations in the track

We showed that the fluctuations in a track represented by SD_{track} are not related to the mean or SD of the SRT estimates for ideal listeners who share the same speech recognition function for the test. Dingemans and Goedegebure (2019) similarly showed that SD_{track} is not useful to detect unreliable measurements. Analyses of experimental data confirmed this finding for adult listeners. This is contrary to the general belief that SD_{track} is a reliable way to detect unreliable results. In many studies, SRTs with large SD_{track} values are classified as unreliable (e.g., Koole *et al.*, 2016; Jacobi *et al.*, 2017; Sheikh Rashid *et al.*, 2017; Denys *et al.*, 2018). Although some unreliable SRT measurements may have large values of SD_{track} , even reliable adaptive tracks

can have values 50% larger than the median values. For such an approach not to incorrectly classify reliable results as unreliable, the criterion for the maximum acceptable SD_{track} value must be set very high. If the test population includes listeners with both normal hearing and hearing impairment, the maximum acceptable SD_{track} value must be near the upper edge of the distribution of SD_{track} values that commonly occur for listeners with the flattest speech recognition functions among the test recipients. In general, these are listeners with the most severe hearing loss. While inattention or lack of understanding of the test can then cause even flatter speech recognition functions in listeners whose results we would like to classify as unreliable, the large spread of SD_{track} values possible for any value of slope means that SD_{track} will not reliably detect those cases. How many it detects depends on how shallow the slope actually was during the test. The risk of using a too strict criterion for the maximum acceptable SD_{track} value is the introduction of a systematic error because the likelihood of excluding test results increases with SRT.

C. Stopping rules

Keidser *et al.* (2013) described an algorithm (or “stopping rule”), proposed by Cameron and Dillon (2007) to control the adaptive procedure in speech-in-noise testing. They calculate a SE after each presentation, and an additional presentation is given until the maximum number of presentations is reached or SE is below a certain criterion. SE is defined as the SD of the presentation levels (i.e., SD_{track} from the current study) divided by the square root of the number of presentations and multiplied by a correction factor [Eq. (1) in Keidser *et al.* (2013)]. Note that the square root is missing for the factor $[1/(n - 1)]$ in their equation for $std(x)$. The proposed algorithm has been used in various papers (e.g., Westermann and Buchholz, 2015; Dillon *et al.*, 2016). The hypothesis is that the SEM for each test is more equal when using a stopping rule than when using a fixed number of presentations. This stopping rule, however, seems to contradict the results of the present study because in our simulations and experimental data, SD_{track} is only weakly related to S_{50} and hence to SEM. We performed additional Monte Carlo simulations to gain more insight into this apparent contradiction. First, the same underlying speech recognition function was used in all the simulations. The results show that the stopping rule introduces differences in track length, but the relationship between SEM and track length is essentially identical to the relationship between SEM and track length for fixed-length procedures. Second, a wide variety of slope values for the underlying speech recognition function were used in the Monte Carlo simulations. Then tracks based on shallow speech recognition functions have, on average, larger values of SD_{track} and, thus, are more likely to fail the criterion of the stopping rule. These tracks will become longer (additional presentations), which reduces the SEM of the associated SRT estimates. It was found that applying the stopping rule to such a heterogeneous group of simulated listeners

yielded a smaller average SEM than for a fixed-length procedure with the same number of presentations as the average number of presentations from the stopping rule procedure. The difference between average SEM values from both procedures, however, was relatively small. The use of various stopping rules in Bayesian adaptive threshold estimation was extensively studied by Alcalá-Quintana and García-Pérez (2005). They reported that none of the stopping rules considered outperformed fixed-length procedures for these Bayesian procedures.

D. Non-ideal listeners

The results from our simulations demonstrate that SD_{track} does not reliably discriminate between poor SRTs due to the non-ideal behavior simulated and SRTs from ideal normal-hearing and hearing-impaired listeners. The characteristics of the underlying population affect the absolute number of rejected SRTs, the percentage of these SRTs that are incorrectly rejected, and the optimal cutoff SD_{track} . If, for example, the percentage of hearing-impaired listeners is very low and the percentage of inattentive listeners is high, a slightly stricter criterion could be used to reject, for example, 50% of the inattentive listeners. Of course, this is at the expense of the rejection of a relatively high percentage of the hearing-impaired listeners and thus still not recommended for most applications.

We did not simulate a typical characteristic noted by Wightman *et al.* (1989) when measuring temporal resolution in children. They reported substantial variation from day to day despite stable performance on each individual measurement: Many of the children seemed to perform at different “levels” on different days (Wightman *et al.*, 1989). It is clearly impossible to identify such an aberrant pattern from the track. Denys *et al.* (2022) administered digits-in-noise tests either as a self-test or as an (adult) administrator-controlled test in a group of normal-hearing children. Our simulations (Fig. 7) seem in line with their experimental results. They did not include children with hearing loss, so a stricter criterion for SD_{track} could be applied to their data. However, applying a cutoff value for SD_{track} of 2.6 dB, corresponding to the 95th percentile for normal-hearing listeners in our simulations (Sec. III C), left a considerable overlap of reliable and unreliable SRTs. In addition, such a strict criterion would exclude too many reliable SRTs from listeners with hearing loss. The individual tracks reported as supplementary material to Denys *et al.* (2022) suggest that at least some of the unreliable SRTs in their study were due to off-route listeners.

The effect of wandering too far away from the SRT estimate (off-route listeners) is strong for relatively short tracks; if the number of presentations is much longer, the effect will become negligible. A likely cause of this type of non-ideal behavior is that the listener does not initially understand the task required of them but then suddenly understands it after several presentations. For the parameters used in our Monte Carlo simulations, a cutoff value of 3.1 dB is sufficient to reject nearly 100% of these cases.

Other measures can be used and may be just as helpful, such as the percentage of correct responses, the number of consecutive correct/incorrect responses, or the number of reversals.

It is important to emphasize that the results of our Monte Carlo simulations only apply to the specific parameters we used in the simulations. The slope and SRT of the normal-hearing speech recognition function and the step size influence the SD_{track} and the pass rate shown in Fig. 7(B). In addition, SRT estimates of listeners with hearing loss can be unreliable as well. We performed additional Monte Carlo simulations, which suggested that using different parameters do not significantly alter the conclusions.

E. Using SD_{track} to estimate the slope of the speech recognition function or to detect unreliable measurements

We have demonstrated that SD_{track} is related to the slope of the speech recognition function. However, it serves as a poor estimator of the true slope. This conclusion is mainly based on Monte Carlo simulations, but also the experimental data from Fig. 5(B) confirm this conclusion. Average SD_{track} values range from 1.4 to 2.7 dB, which correspond to slope values from $>25\%/dB$ to $4\%/dB$ [Fig. 6(A)], although the average SRT and underlying speech recognition function are the same. Our results are in line with expectations because the adaptive procedure ensures that the SNRs of the presentations are close to the SRT. It has been previously indicated in seminal papers that it is important to include presentation levels sufficiently far from the SRT when estimating the slope (Dixon and Mood, 1948; Levitt, 1971). There is a lot of literature on possible estimates of both the SRT and the slope. Without striving for completeness, we mention the use of two transformed adaptive procedures (Levitt, 1971): the procedure proposed by Brand and Kollmeier (2002) and a Bayesian adaptive estimation (Doire *et al.*, 2017). An extensive overview of different procedures can be found in Doire *et al.* (2017).

Our results show that measures other than SD_{track} are required to determine the reliability of a single measurement. Several listener-related factors could negatively affect the accuracy of a measurement. These factors are, among others, a training or learning effect and lapses (unforced errors) due to fatigue, inattentiveness, misunderstanding of the response by the experimenter, or incorrectly entering the response. These factors are difficult to control, and it is often unclear whether they are present or not. Because inattention is probably a leading cause of lapses, a measure of attention could be a valuable addition. Moore *et al.* (2010) studied auditory processing disorder in a large group of normal-hearing children. They reported that intrinsic measures of attention, derived from tracking variable responses, did relate in general to non-speech auditory performance, but not to speech-in-noise [vowel-consonant-vowel (VCV)] recognition. Similarly, Denys *et al.* (2022) did not find significant differences in attentional abilities between groups of children who had stable SRTs and those who showed

differences between repeatedly measured SRTs. They also reported that SRTs from children with poorer attention skills were not significantly worse than SRTs from children with better attention skills. Different forms of attentional abilities were measured: selective attention, sustained attention, and attention switching. In addition, a teacher questionnaire was used. On the other hand, Riccio *et al.* (2005) reported a moderate correlation between a measure of sustained attention and speech-in-noise test scores.

We are currently considering two measures that could potentially be used to identify unreliable measurements. First, we consider response time. We hypothesize that inattentiveness may be reflected in longer response times to stimuli or may lead to erratic response time patterns during a track. Second are relationships between recognition probabilities. On average, due to the nature of the adaptive procedure, 50% of the presentations are recognized correctly and 50% incorrectly. For digits-in-noise tests with three independent digits per presentation, the probability to recognize zero, one, two, or three (i.e., the entire digit triplet) digits correctly should follow a binomial distribution. It seems likely that, for inattentive listeners, these relationships are different. For example, if they easily give up when they cannot understand all the digits, then more often a response with none of the digits recognized correctly will be expected. The value of these measures in identifying unreliable measurements is still highly speculative. Studies are needed to investigate these and other potential measures.

V. CONCLUSIONS

The one-up one-down adaptive (staircase) procedure is a relatively simple and accurate procedure to estimate the SRT in speech-in-noise testing. It has been one of the most popular psychophysical methods for decades. Advantages are the insensitivity to lapses, the lack of necessity to know the exact shape of the speech recognition function, and the small effect of the choice of step size on the SRT estimate. Original findings from early studies that are sometimes overlooked are the following:

- The SEM is inversely proportional to the slope of the speech recognition function and square root of the number of presentations;
- A within-subjects (repeated measures) design should be used to determine the SEM of a test accurately.

The major findings of the current study can be summarized as follows:

- The fluctuations in an adaptive track (represented by SD_{track}) do not provide information about the SEM or how near the SRT estimate is to the true SRT for (near) ideal listeners with the same underlying speech recognition function;
- Hearing loss and non-ideal behavior (inattentiveness, fatigue, and giving up when the task becomes too difficult) slightly increase the average value of SD_{track} .

SD_{track} , however, poorly discriminates between reliable and unreliable SRT estimates;

- Large fluctuations in the track and large systematic errors in SRT estimates are found for listeners who unintentionally respond incorrectly to a few consecutive presentations and need a substantial number of presentations to reach SNRs near the true SRT. Although values of SD_{track} will increase for these listeners, there is still a considerable overlap with SD_{track} values from reliable SRT estimates.

ACKNOWLEDGMENT

D.R.M. and H.D. receive support from the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre.

¹See supplementary material at <https://www.scitation.org/doi/suppl/10.1121/10.0014898> for a comparison between the low SD_{track} group and the high SD_{track} group constructed from test and retest data from Smits *et al.* (2016).

Alcalá-Quintana, R., and García-Pérez, M. A. (2005). “Stopping rules in Bayesian adaptive threshold estimation,” *Spat. Vis.* **18**, 347–374.

Bode, D., and Carhart, R. (1973). “Measurement of articulation functions using adaptive test procedures,” *IEEE Trans. Audio Electroacoust.* **21**, 196–201.

Brand, T., and Kollmeier, B. (2002). “Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests,” *J. Acoust. Soc. Am.* **111**, 2801–2810.

Brownlee, K., Hodges, J., Jr., and Rosenblatt, M. (1953). “The up-and-down method with small samples,” *J. Am. Stat. Assoc.* **48**, 262–277.

Cameron, S., and Dillon, H. (2007). “Development of the listening in spatialized noise-sentences test (LISN-S),” *Ear Hear.* **28**, 196–211.

Denys, S., Hofmann, M., Luts, H., Guérin, C., Keymeulen, A., Van Hoeck, K., van Wieringen, A., Hoppenbrouwers, K., and Wouters, J. (2018). “School-age hearing screening based on speech-in-noise perception using the digit triplet test,” *Ear Hear.* **39**, 1104–1115.

Denys, S., Wouters, J., and van Wieringen, A. (2022). “The digit triplet test as a self-test for hearing screening at the age of school-entry,” *Int. J. Audiol.* **61**, 408–415.

de Vet, H. C., Terwee, C. B., Knol, D. L., and Bouter, L. M. (2006). “When to use agreement versus reliability measures,” *J. Clin. Epidemiol.* **59**, 1033–1039.

Dillon, H. (1983). “The effect of test difficulty on the sensitivity of speech discrimination tests,” *J. Acoust. Soc. Am.* **73**, 336–344.

Dillon, H., Beach, E. F., Seymour, J., Carter, L., and Golding, M. (2016). “Development of Telscreen: A telephone-based speech-in-noise hearing screening test with a novel masking noise and scoring procedure,” *Int. J. Audiol.* **55**, 463–471.

Dingemans, G., and Goedegebure, A. (2019). “Efficient adaptive speech reception threshold measurements using stochastic approximation algorithms,” *Trends Hear.* **23**, 2331216520919199.

Dixon, W. J., and Mood, A. M. (1948). “A method for obtaining and analyzing sensitivity data,” *J. Am. Stat. Assoc.* **43**, 109–126.

Doire, C. S., Brookes, M., and Naylor, P. A. (2017). “Robust and efficient Bayesian adaptive psychometric function estimation,” *J. Acoust. Soc. Am.* **141**, 2501–2512.

García-Pérez, M. A. (1998). “Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties,” *Vision Res.* **38**, 1861–1881.

Green, D. M. (1995). “Maximum-likelihood procedures and the inattentive observer,” *J. Acoust. Soc. Am.* **97**, 3749–3760.

Jacobi, I., Sheikh Rashid, M., de Laat, J., and Dreschler, W. A. (2017). “Age dependence of thresholds for speech in noise in normal-hearing adolescents,” *Trends Hear.* **21**, 2331216517743641.

Keidser, G., Dillon, H., Mejia, J., and Nguyen, C.-V. (2013). “An algorithm that administers adaptive speech-in-noise testing to a specified reliability

- at selectable points on the psychometric function," *Int. J. Audiol.* **52**, 795–800.
- Koole, A., Nagtegaal, A. P., Homans, N. C., Hofman, A., Baatnburg de Jong, R. J., and Goedegebure, A. (2016). "Using the digits-in-noise test to estimate age-related hearing loss," *Ear Hear.* **37**, 508–513.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**(Suppl. 2), 467–477.
- Levitt, H., and Rabiner, L. R. (1967). "Use of a sequential strategy in intelligibility testing," *J. Acoust. Soc. Am.* **42**, 609–612.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., and de Vet, H. C. (2010). "The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study," *Qual. Life Res.* **19**, 539–549.
- Moore, D. R., Ferguson, M. A., Edmondson-Jones, A. M., Ratib, S., and Riley, A. (2010). "Nature of auditory processing disorder in children," *Pediatrics* **126**, e382–e390.
- Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.* **95**, 1085–1099.
- Plomp, R., and Mimpen, A. M. (1979). "Improving the reliability of testing the speech reception threshold for sentences," *Int. J. Audiol.* **18**, 43–52.
- Riccio, C. A., Cohen, M. J., Garrison, T., and Smith, B. (2005). "Auditory processing measures: Correlation with neuropsychological measures of attention, memory, and behavior," *Child Neuropsychol.* **11**, 363–372.
- Sheikh Rashid, M., Leensen, M. C. J., de Laat, J., and Dreschler, W. A. (2017). "Laboratory evaluation of an optimised internet-based speech-in-noise test for occupational high-frequency hearing loss screening: Occupational earcheck," *Int. J. Audiol.* **56**, 844–853.
- Smits, C., and Festen, J. M. (2011). "The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: Steady-state noise," *J. Acoust. Soc. Am.* **130**, 2987–2998.
- Smits, C., and Houtgast, T. (2005). "Results from the Dutch speech-in-noise screening test by telephone," *Ear Hear.* **26**, 89–95.
- Smits, C., and Houtgast, T. (2006). "Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests," *J. Acoust. Soc. Am.* **120**, 1608–1621.
- Smits, C., Kapteyn, T. S., and Houtgast, T. (2004). "Development and validation of an automatic speech-in-noise screening test by telephone," *Int. J. Audiol.* **43**, 15–28.
- Smits, C., Theo Goverts, S., and Festen, J. M. (2013). "The digits-in-noise test: Assessing auditory speech recognition abilities in noise," *J. Acoust. Soc. Am.* **133**, 1693–1706.
- Smits, C., Watson, C. S., Kidd, G. R., Moore, D. R., and Goverts, S. T. (2016). "A comparison between the Dutch and American-English digits-in-noise (DIN) tests in normal-hearing listeners," *Int. J. Audiol.* **55**, 358–365.
- Taylor, M. (1971). "On the efficiency of psychophysical measurement," *J. Acoust. Soc. Am.* **49**, 505–508.
- Theunissen, M., Swanepoel, D. W., and Hanekom, J. (2009). "Sentence recognition in noise: Variables in compilation and interpretation of tests," *Int. J. Audiol.* **48**, 743–757.
- UK Biobank (2012). "UK Biobank Hearing 'Speech-in-Noise' Test version 1.3," <https://biobank.ndph.ox.ac.uk/ukb/docs/Hearing.pdf> (Last viewed October 18, 2022).
- Westermann, A., and Buchholz, J. M. (2015). "The influence of informational masking in reverberant, multi-talker environments," *J. Acoust. Soc. Am.* **138**, 584–593.
- Wetherill, G. (1963). "Sequential estimation of quantal response curves," *J. R. Statistical Soc. Ser. B* **25**, 1–48.
- Wetherill, G., Chen, H., and Vasudeva, R. (1966). "Sequential estimation of quantal response curves: A new method of estimation," *Biometrika* **53**, 439–454.
- Wetherill, G., and Levitt, H. (1965). "Sequential estimation of points on a psychometric function," *Br. J. Math. Stat. Psychol.* **18**, 1–10.
- Wightman, F., Allen, P., Dolan, T., Kistler, D., and Jamieson, D. (1989). "Temporal resolution in children," *Child Dev.* **60**, 611–624.
- Zokoll, M. A., Wagener, K. C., Brand, T., Buschermöhle, M., and Kollmeier, B. (2012). "Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization prototype," *Int. J. Audiol.* **51**, 697–707.