# Development of an adaptive scaling method for subjective listening effort[a)]

Melanie Krueger[b),c)] and Michael Schulte[c)]
*Hörzentrum Oldenburg, Marie-Curie-Straße 2, D-26129 Oldenburg, Germany*

Thomas Brand[c)]
*Medizinische Physik, Department für Medizinische Physik und Akustik, Fakultät VI, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany*

Inga Holube[c)]
*Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Ofener Straße 16/19, D-26121 Oldenburg, Germany*

(Received 21 December 2016; revised 7 June 2017; accepted 7 June 2017; published online 22 June 2017)

An adaptive procedure for controlling the signal-to-noise ratio (SNR) when rating the subjectively perceived listening effort (Adaptive Categorical Listening Effort Scaling) is described. For this, the listening effort is rated on a categorical scale with 14 steps after the presentation of three sentences in a background masker. In a first phase of the procedure, the individual SNR range for ratings from "no effort" to "extreme effort" is estimated. In the following phases, stimuli with randomly selected SNRs within this range are presented. One or two linear regression lines are fitted to the data describing subjective listening effort as a function of SNR. The results of the adaptive procedure are independent of the initial SNR. Although a static procedure using fixed, predefined SNRs produced similar results, the adaptive procedure avoided lengthy pretests for suitable SNRs and limited possible bias in the rating procedures. The adaptive procedure resolves individual differences, as well as differences between maskers. Inter-individual standard deviations are about three times as large as intra-individual standard deviations and the intra-class correlation coefficient for test-retest reliability is, on average, 0.9. © *2017 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4986938]

[AKL] Pages: 4680–4693

## I. INTRODUCTION

In recent years, the effort involved in listening to speech has received increasing attention. Many listeners with hearing impairment perceive noisy environments as tiring and describe understanding speech as effortful (Pichora-Fuller *et al.*, 2016). Individually differing amounts of mental processing resources are involved in listening under adverse conditions (Lemke and Besser, 2016) and physiological stress responses can be evoked (Mackersie and Calderon-Moultrie, 2016). These phenomena were addressed by different authors by measuring the listening effort. However, different methods and definitions like "ease of listening," "listening effort," or "listening difficulty" were used that did not necessarily lead to the same results. A white paper issued by the Cognition in Hearing Special Interest Group of the British Society of Audiology (McGarrigle *et al.*, 2014) defined listening effort as "the mental exertion required to attend to, and understand, an auditory message." In addition, an Eriksholm workshop on Hearing Impairment and Cognitive Energy proposed defining listening effort more generically as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task" (Pichora-Fuller *et al.*, 2016).

Listening effort can be measured using objective or subjective methods (see Klink *et al.*, 2012a,b, for an overview). Objective methods are single- or dual-task paradigms or physiological measurements. Subjective methods include questionnaires and rating scales, e.g., the standardized and validated "Speech, Spatial and Qualities of Hearing Scale (SSQ)" (Gatehouse and Noble, 2004), the "NASA Task Load Index" (NASA-TLX; Hart and Staveland, 1988), or a categorical scale (Schulte *et al.*, 2009; Luts *et al.*, 2010).

In this article, the term "listening effort" is used solely for the subjectively perceived effort that is required to follow a few sentences of one speaker in background noise. This approach was used although listener ratings might be influenced by the speech intelligibility, or the task difficulty and the amount of mental resources required to perform the speech test might remain hidden (McGarrigle *et al.*, 2014). However, Johnson *et al.* (2015) and Holube *et al.* (2016) showed that the assessment of subjective listening effort using a categorical scale is practicable, quick, and easy to administer. The applied subjective method was also superior over a word recall measure and a physiological skin conductance measure because of its ability to differentiate diverse (i.e., easier and more difficult) listening situations.

[a)] Parts of the study were presented at the annual meeting of the Deutsche Gesellschaft für Audiologie (DGA, 2015) in Bochum, Germany.
[b)] Electronic mail: m.krueger@hoerzentrum-oldenburg.de
[c)] Also at: Cluster of Excellence "Hearing4All," Carl-von-Ossietzky University Oldenburg, D-26111 Oldenburg, Germany.

0001-4966/2017/141(6)/4680/14/$30.00

In previous studies, subjective listening effort was measured in listening situations with pre-defined conditions, e.g., several fixed signal-to-noise ratios (SNRs) (Sato *et al.*, 2005; Zekveld *et al.*, 2011; Luts *et al.*, 2010; van Schoonhoven *et al.*, 2016). It is known that the range and the order of stimuli influence subjective ratings, and that in order to avoid biases, stimuli should be evenly distributed and randomly presented in the individual range of perception (Parducci and Perrett, 1971; Montgomery, 1975; Poulton, 1989; Zielinski, 2016). In the present case, the limits of the individual perception range can be attributed to "no effort" and "extreme effort." The number of stimuli *at* those limits, which do not contribute to modeling subjective listening effort in dependence on the SNR *within* the limits, should be small, to restrict the measurement duration. Therefore, in order to realize a proper stimulus placement using fixed predefined stimuli, a premeasurement is typically required.

In the present study, a procedure for Adaptive Categorical Listening Effort Scaling (ACALES) is introduced that aims for proper stimulus placement by including premeasurements directly within the procedure. The procedure is based on the adaptive procedure for categorical loudness scaling (ACALOS; Brand and Hohmann, 2002). ACALOS quantifies loudness perception for different levels ranging from "inaudible" to "too loud." ACALOS controls each presentation level based on previous ratings of perceived loudness. Accordingly, ACALES controls each presentation SNR based on previous ratings of perceived listening effort. On the basis of the listener's ratings in the first trials of the measurement, ACALES iteratively adapts to the parameter range. At the beginning of the procedure, the boundaries of the parameter range (SNR values in this study) for ratings of "no effort" and "extreme effort" are estimated. In the subsequent measurement phase, stimuli are distributed evenly in the parameter range thus estimated and are presented in randomized order. Only the initial SNR has to be set.

Various scales have been proposed to rate listening effort. They differ in the number of categories and the corresponding notations. A 13-point scale was used by Luts *et al.* (2010). The scale ranges from "no effort" to "extreme effort" and was also applied by, e.g., Rennies *et al.* (2014), Schepker *et al.* (2016), and Holube *et al.* (2016) and is also used for ACALES. When using this or similar scales, a particular challenge for the measurement protocol is the selection of relevant test conditions. Larsby *et al.* (2005) observed that the perceived effort deviated for different background noises and for quiet. Furthermore, they observed increased effort for hearing-impaired (HI) listeners compared to normal-hearing (NH) listeners. Rating scores might also manifest individual differences (McGarrigle *et al.*, 2014) determined by an individual's experiences and expectations. Larsby *et al.* (2005) speculated, for example, that the observed difference between HI and NH listeners might at least in part be related to the unfamiliar listening condition without hearing aids for the hearing-impaired listeners. Before applying the new ACALES procedure to different subject groups, and before exploring the potential easing in subjective listening effort with hearing aids, the procedure itself has to be examined. Therefore, this study describes the ACALES procedure and its evaluation using NH listeners. The main questions are:

(1) Does the initial SNR influence the result? The hypothesis is that the initial SNR is used as an anchor and that listeners' responses tend to rate it with medium effort.
(2) Are the results of the adaptive procedure different from the procedure using fixed static SNRs comparing the SNRs of the rating categories? This constant stimuli procedure is regarded as the gold standard.
(3) Can ACALES resolve differences of perceived listening effort between the NH listeners used in this study? Differences can be resolved if the intra-individual standard deviation of results is smaller than the inter-individual standard deviation.
(4) Can ACALES resolve differences of perceived listening effort between different background noises? The hypothesis is that the perceived listening effort is not only determined by the SNR but that the noise type causes different ratings that can be significantly resolved by the procedure.
(5) Are the results of the adaptive procedure reliable? The hypothesis is that the listeners are able to reproduce their listening effort ratings in test-retest conditions.

## II. METHODS

The research questions formulated in Sec. I were addressed in two experiments described in the following sections. Table I gives an overview of the participants, the maskers and the objectives of the experiments.

### A. Participants

All 25 participants had normal hearing, defined as a pure tone average (PTA$_4$; average of 500 Hz, 1, 2, and 4 kHz) of less than 20 dB hearing level.

TABLE I. Participants, maskers and objectives examined in the two experiments.

| | Participants | Stationary maskers | Fluctuating maskers | Objectives |
|---|---|---|---|---|
| Experiment 1 | 10 NH listeners | • Olnoise | • ISTS | A. Determination of fit function<br>B. Effect of initial SNR<br>C. Static vs. adaptive procedure |
| Experiment 2 | 15 NH listeners | • Olnoise<br>• Cafeteria | • IFFM<br>• Icra5-250 | D. Masker comparison<br>E. Intra- and inter-individual standard deviation<br>F. Test-retest reliability |

### 1. Participants for experiment 1

For experiment 1, ten subjects aged between 19 and 31 years (mean age: 23.8 years; male/female: 2/8) were invited to the Hörzentrum Oldenburg for one session. The mean PTA$_4$ of this group was 0.7 dB for the left [standard deviation (SD): 2.0 dB] and 1.5 dB for the right ear (SD: 3.1 dB). Five of the ten subjects had no experience in the field of speech tests and listening effort ratings and were students from different faculties of Carl von Ossietzky University in Oldenburg. The other five subjects were students of the "Hearing Technology and Audiology" study program in Oldenburg and were familiar with the Oldenburg sentence test (OLSA), but not with the listening effort ratings.

### 2. Participants for experiment 2

This group consisted of 15 listeners aged between 21 and 31 years (mean age: 24.6 years; male/female: 9/6). They were invited to the Hörzentrum Oldenburg for three sessions. The mean PTA$_4$ of these subjects was 1.9 dB for the left (SD: 4.5 dB) and 2.3 dB for the right ear (SD: 3.1 dB). The subjects were recruited from different faculties of Carl von Ossietzky University in Oldenburg and had no experience with listening effort rating measurements.

## B. Stimuli

### 1. General

Subjects listened to three different sentences of the OLSA (Wagener *et al.*, 1999a,b; Wagener *et al.*, 1999c) before each rating of their subjective listening effort. All OLSA sentences were spoken by one male speaker and consisted of five words belonging to a predefined word class (name – verb – numeral – adjective – object) and presented in the same order. One example is "Nina malt zehn nasse Sessel" ("Nina paints ten wet armchairs" in English). For each of the five word classes, ten possible word alternatives existed. A presentation of three sentences was chosen to allow for a reasonable amount of time to listen to the stimuli and to decide how effortful the listening was.

The sentences were presented in randomly chosen time segments of different maskers. The presentation of the maskers started 2 s before the presentation of the first sentence and was switched off at the end of the third and last sentence. Hence, the masker was not present during the response time of the subjects but started again within 1 s after the response of the subjects. The average duration of one sentence was 2.2 s (SD: 0.2 s) and the time interval between the sentences was 0.7 s, resulting in a total presentation time of 10 s for the masker.

All maskers were calibrated to a level of 65 dB sound pressure level (SPL). Since the sentences were presented at several SNRs (see descriptions on the procedures below), the overall level of masker and sentence was 65 dB SPL or higher, depending on the selected SNR.

### 2. Maskers in experiment 1

In experiment 1, the OLSA sentences were presented in two different maskers: the fluctuating International Speech Test Signal (ISTS; Holube *et al.*, 2010) and the stationary Olnoise (Wagener *et al.*, 1999a,b; Wagener *et al.*, 1999c). The ISTS was based on recordings of six different female speakers reading the fable of "North Wind and Sun" in their mother tongue (American English, Arabic, Mandarin, French, German, and Spanish). Because of segmentation and mixing of the six different recordings, the signal was mostly incomprehensible, but the temporal and spectral characteristics were similar to the characteristics of one single female speaker. The Olnoise was generated from the speech stimuli used in the OLSA (for more details see Wagener *et al.*, 1999a). As a result, the Olnoise provides the same long-term average spectrum as the speech of the OLSA and has a nearly stationary temporal characteristic.

### 3. Maskers in experiment 2

In experiment 2, the OLSA sentences were presented in two stationary and two fluctuating maskers. The two stationary maskers were Olnoise and Cafeteria noise. The Cafeteria noise was recorded in the cafeteria of Carl von Ossietzky University in Oldenburg. The two fluctuating maskers were International Female Fluctuating Masker (IFFM) and Icra5-250. For IFFM, the ISTS was modified by shortening the maximal pause durations to 250 ms (Holube, 2011). For Icra5-250, the Icra5 signal, a noise signal with an envelope characteristics of one male speaker (Dreschler *et al.*, 2001), was also modified by shortening the maximal pause durations to 250 ms (Wagener *et al.*, 2006).

## C. Categorical rating scale

Subjective listening effort was measured by asking "How much effort does it require for you to follow the speaker?" ("Wie anstrengend ist es für Sie, dem Sprecher zu folgen?" in German). This phrasing commonly also includes understanding the words themselves, especially as the content of the sentences was easy to understand. The responses were given on a categorical rating scale with seven labeled categories and six intermediate steps, as used by Luts *et al.* (2010) and complemented by the category "only noise" ("nur Störgeräusch" in German; see Fig. 1 for the original screenshot of the German version and the English translation). The category "only noise" was introduced to allow for a response when no speech cues were perceived. Accordingly, this response was not related to any perceived listening effort. Pretests including a comparison of the scale with and without the additional category "only noise" revealed that subjects described the additional category as helpful. When "only noise" was not included in the scale, subjects rated situations with undetectable speech as "extremely effortful" not allowing for a differentiation from situations with low speech levels, which are really extremely effortful to listen to.

Effort scale categorical units (ESCU) were assigned to the categories as numerical entities. The category "no effort" ("mühelos" in German) corresponded to 1 ESCU, "very little

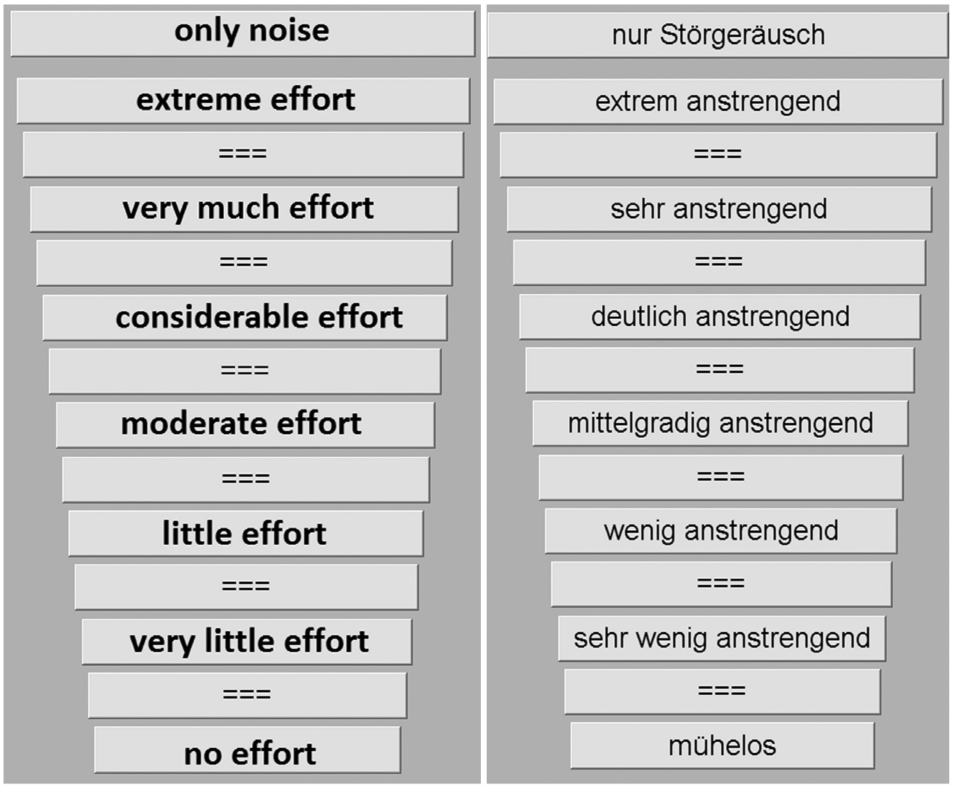

FIG. 1. Rating scale for listening effort in English (left) and the original version in German (right).

effort" ("sehr wenig anstrengend") to 3 ESCUs, "little effort" ("wenig anstrengend") to 5 ESCUs, "moderate effort" ("mittelgradig anstrengend") to 7 ESCUs, "considerable effort" ("deutlich anstrengend") to 9 ESCUs, "very much effort" ("sehr anstrengend") to 11 ESCUs, "extreme effort" ("extrem anstrengend") to 13 ESCUs. The numbers in ESCUs were not visible to the subjects. The listening effort was rated for a predefined range of SNRs (static procedure) or for an individually adjusted range of SNRs (adaptive procedure).

### D. Adaptive procedure

The adaptive procedure for rating the listening effort (ACALES) was based on the adaptive procedure for categorical loudness scaling (ACALOS) of Brand and Hohmann (2002). Within these procedures, the term "adaptive" does not describe an adaptive variation of one presentation level as carried out, for instance, in a speech test to determine the Speech Reception Threshold (SRT). As in the ACALOS procedure, the operative range of parameter settings, levels in ACALOS, and SNRs in ACALES, was changed adaptively in ACALES. The ACALES procedure was divided into three phases:

In the first phase, the boundaries of the SNR range for the ratings "no effort" and "extreme effort" or "only noise" were determined (see Fig. 2 for an example). The SNR was adaptively modified based on the ratings of the subjects in two interleaved search processes, which were run in alternate order. In one search process, the SNR was increased in 3 dB steps until the subjects rated the presentation as "no effort." In the other search process, the level was decreased in 3 dB steps until "extreme effort" or "only noise" was selected. This phase ended when the boundaries 1 and 13 or 14 of the rating scale were found. In the example shown in Fig. 2, this was the case at −12 dB and 9 dB. The step size was chosen to be 3 dB, as this corresponds to the Just-Noticeable-Difference for SNRs (McShefferty *et al.*, 2015). If the subject rated the first presentation as "no effort" or "extreme effort"/"only noise," the SNR was modified with a step size of 5 dB until the subject selected a response between 2 and 12 ESCU.

The boundaries were then used in the second phase of the procedure to estimate the SNRs for the five categories "very little effort," "little effort," "moderate effort," "considerable effort," and "very much effort" by linear interpolation. These five SNRs were presented once to the subjects in random order. The subjects were not aware of the intended targets for listening effort rating but rated their

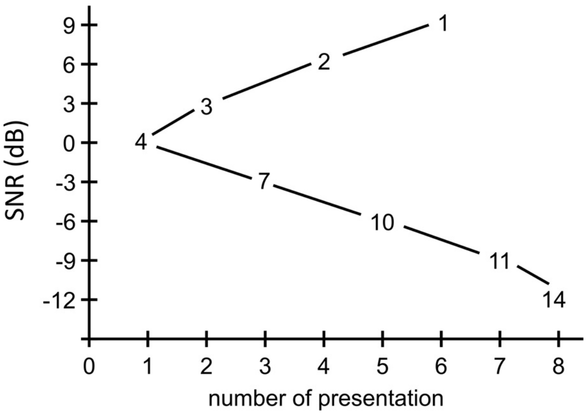

FIG. 2. Example of the first phase of the adaptive method to determine the boundaries of the rating scale ("no effort" and "extreme effort"/"only noise"). The numbers in the figure indicate the rated listening effort in ESCU for each presentation of three sentences, respectively.

TABLE II. Selected SNRs for the static method using Olnoise and ISTS. The second column gives the SNRs for the first presentation, and the number of subjects in parentheses for whom this SNR was selected. The subsequent presentations were selected randomly from the values given in the respective rows.

| | First SNR (dB) | Subsequent SNRs (dB) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Olnoise | −5 (3), 0 (3),<br>5 (2), 10 (2) | −10 | −7 | −4 | −1 | 2 | 5 | 8 |
| ISTS | −20 (3), −10 (3),<br>0 (2), 10 (2) | −23 | −17 | −11 | −5 | 1 | 7 | — |

subjective listening effort on the scale described in Sec. II C based on the presented stimuli.

In the third phase, a straight line was fit to the ratings of the second phase by linear regression and the SNRs for "no effort" and "extreme effort" were re-estimated. On the basis the new boundaries, SNRs for the five categories "very little effort," "little effort," "moderate effort," "considerable effort," "very much effort" were re-calculated and an SNR for the sixth category "extreme effort" was added and presented to the listeners. Re-estimation of the boundaries and re-calculation of the six SNRs was repeated, resulting in two presentations for each of the six categories within the third phase. To avoid too-soft or too-loud sound presentations, the minimal speech level was set to 40 dB SPL (−25 dB SNR) and the maximal speech level to 95 dB SPL (30 dB SNR). Those minimal and maximal SNRs thus replaced presentations of estimated higher or lower values.

## E. Static procedure

Within the static method, the first presentation was either at −5, 0, 5, or 10 dB SNR for Olnoise (see Table II). For ISTS, presentations at −20, −10, 0, or 10 dB SNR were used. These initial SNRs were selected to be comparable to the initial SNRs for the adaptive procedure (see Table III). The SNRs for the subsequent presentations were set between −10 and 8 dB SNR with 3 dB difference for Olnoise, or between −23 and 7 dB SNR with 6 dB difference for ISTS. Each SNR given in columns 3−9 in Table II was presented three times. The order was randomized.

The SNR range was selected with the intention of covering the whole range between "no effort" and "extreme effort." The range for Olnoise was chosen based on the results of Schulte *et al.* (2009) and Luts *et al.* (2010). Pretests with normal-hearing listeners not included in the further analysis revealed that the perceived listening effort depended on the masker type, with a shift to lower SNRs and a larger SNR range for ISTS relative to Olnoise. Therefore, the step size between the SNRs was 6 dB for ISTS compared to 3 dB for Olnoise.

TABLE III. Selected initial SNRs for the adaptive method using Olnoise and ISTS. The level of the masker was always 65 dB SPL.

| | Initial SNRs (dB) | | | |
|---|---|---|---|---|
| Olnoise | −5 | 0 | 5 | 10 |
| ISTS | −20 | −10 | 0 | 10 |

## F. Experiments

Prior to data collection, all participants received information about the study and consent was obtained. After measuring the pure tone audiogram (250 Hz to 8 kHz), listening effort was rated under different conditions.

At first, training was performed to demonstrate the test procedure and to familiarize the subjects with the maskers and the rating scale. The training consisted of the first five SNRs of an adaptive procedure for each masker. The initial SNR for training was always 0 dB SNR and the following four SNRs were selected as shown in Fig. 2: 3, −3, 6, and −6 dB SNR. As in the following measurements, the sentences were selected randomly from the set of sentences in the Oldenburg speech test.

### *1. Experiment 1*

In experiment 1, the static and the adaptive procedure were used. For the adaptive method, different initial SNRs were selected (see Table III), to evaluate their influence on the outcome. The initial SNRs were chosen in the center as well as on the boundaries of the SNR ranges for Olnoise and ISTS used in the static procedure (see Table II), to examine possible floor or ceiling effects.

Thus each subject completed ten runs during one session. A run consisted of one adaptive procedure as described in Sec. II D or one static procedure as described in Sec. II E. The ten runs were divided into one run for each of the two maskers using the static method and four runs for each of the two maskers using the adaptive method. The four runs using the adaptive method differed in their initial SNR. To avoid training effects, the order of conditions was randomized (static vs adaptive method, Olnoise vs ISTS, and initial SNRs). The subjects were not informed about the different conditions and identical written instructions were used.

### *2. Experiment 2*

In experiment 2, only the adaptive procedure was used. Subjects were appointed for three visits to evaluate test-retest reliability. In each session, the subjects rated the perceived listening effort for all four maskers, whereby the presentation order was randomized. The initial SNR was always 0 dB SNR.

## G. Equipment

The measurements were performed in a soundproof room at the Hörzentrum Oldenburg. All signals were D/A converted (sound card ADI-8 Pro by RME, Haimhausen, Germany), amplified (HB7 by Tucker-Davis, Alachua, FL) and presented to the subjects from the frontal direction via a loudspeaker (Mackie HR 824 by LOUD technologies, Woodinville, WA). The subjects were seated at a distance of 1.4 m from the loudspeaker; the head position was not fixed. For the graphical presentation of the response scale and the response input by the subjects themselves, a touch screen was used. The level of the Olnoise was calibrated to 65 dB SPL with the listener absent, using a measurement microphone (type 4189 by Brüel and Kjær, Nærum, Denmark) at the position of the listener, and a sound level meter ("Modular Precision Sound Analyzer"; model 2260 by Brüel and Kjær, Nærum, Denmark).

### H. Analysis and statistics

In a first step, the individual results of each subject for rated listening effort as a function of the SNR were analyzed. Example results of two subjects in two measurement conditions using the adaptive procedure are shown in Fig. 3. The BX fitting method developed by Oetting *et al.* (2014) was used to fit a function to the individual data points, but not including the rating category "only noise" (o.n.). This method fits a function to the data points by minimizing the deviation of the data points on the SNR axis and not on the categorical response axis. The BX fitting method applied a linear regression line for the categories 1 to 7 ESCU and another regression line including the categories 7 to 13 ESCU, resulting in a two-slope function. The crossing point is smoothed between the categories 5 and 9 ESCU with Bezier transition. This two-slope model function was selected because visual inspection showed that it allowed for minimal deviation from both, mean and median SNRs of all subjects for each effort category under every condition. For Olnoise, the resulting curve is very similar to a linear regression line, as the slopes of the upper and the lower part of the curve are almost identical (see Sec. III). For the fluctuating masker ISTS, the model function consisting of two regression lines with different slopes approximated the data much better than a single linear regression line and was therefore applied to all maskers. The two-slope model function enables the description of measurement results with three parameters (upper slope, crossing point, and lower slope) and facilitates comparisons between subjects and measurement conditions. Mean two-slope functions for each measurement condition were calculated by using all data points of the individual listeners in the respective condition.

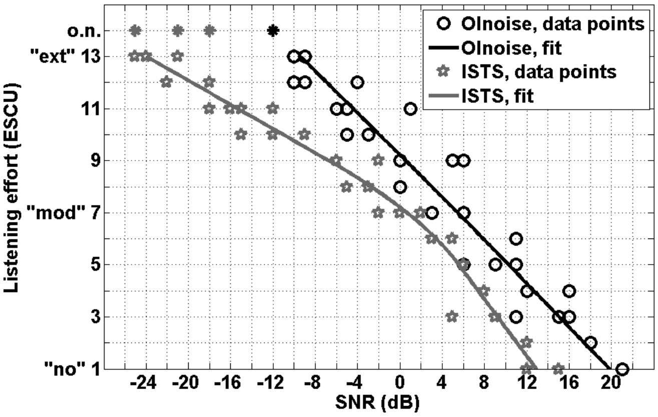

FIG. 3. Two examples for listening effort data for sentences presented in Olnoise and ISTS at different SNRs. Model functions composed of two linear regression lines were fitted to the data points.

To validate the quality of the fit, $R_p{}^2$ of the prediction was calculated using the ratio between the predicted residual error sum of squares and the total sum of squares:

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left(\mathrm{SNR}_i - \widehat{\mathrm{SNR}}_{i,-i}\right)^2}{\sum_{i=1}^{n}\left(\mathrm{SNR}_i - \overline{\mathrm{SNR}}\right)^2}, \qquad (1)$$

$\mathrm{SNR}_i$ is the observed SNR and $\widehat{\mathrm{SNR}}_{i,-i}$ the value, which can be estimated if all observations except $i$ itself were included in the model.

To compare different conditions (different initial SNRs for the adaptive procedure, static vs adaptive procedure), the SNR was calculated for seven rating categories ("no effort," "very little effort," "little effort," "moderate effort," "considerable effort," "very much effort," and "extreme effort") for each subject and condition using the BX fitting method with the two-slope function. After testing for normal distribution of the data, an analysis of variance (ANOVA) for repeated measurements was performed for all calculated SNRs. Data were analyzed with the software SPSS, Version 22.

## III. RESULTS

The individual fits, the impact of the initial SNR, and a comparison between the static and the adaptive procedure were analyzed using the data of experiment 1 (Secs. III A–III C), whereas differences between masker types, intra- and inter-individual differences as well as the test-retest reliability were analyzed using the data of experiment 2 (Secs. III D–III F).

### A. Individual fits

Figure 4 shows fitted two-slope functions for perceived listening effort in experiment 1 using the adaptive procedure and the maskers Olnoise and ISTS. For all subjects but two and for each estimate, the quality parameter $R_p{}^2$ was between 0.868 and 0.948, with a mean of 0.878 in Olnoise and a mean of 0.893 in ISTS. For the two remaining subjects, $R_p{}^2$ was 0.776 and 0.710 in Olnoise and between 0.801 and 0.838 in ISTS. These results indicate that the BX fitting method using the two-slope function approximated the data well and was applicable to Olnoise as well as to ISTS.

The observed two-slope functions revealed large inter-individual differences. The SNRs for "extreme effort" ranged from −11 to −3 dB SNR for the Olnoise and from −25 to −15 dB SNR for the ISTS. Even larger SNR ranges were observed for the rating category "no effort" (3 to 19 dB SNR for Olnoise and 2 to 27 dB SNR for ISTS).

To simplify the display of different conditions, overall estimated two-slope functions were calculated using the data of all subjects in the respective condition as input for the BX fitting method (see Fig. 5). The resulting overall estimated two-slope functions for Olnoise and ISTS are also shown as black lines in Fig. 4. The overall estimates are a good approximation to the mean SNR values in each rating category shown as a dashed line in Fig. 5. The advantage of

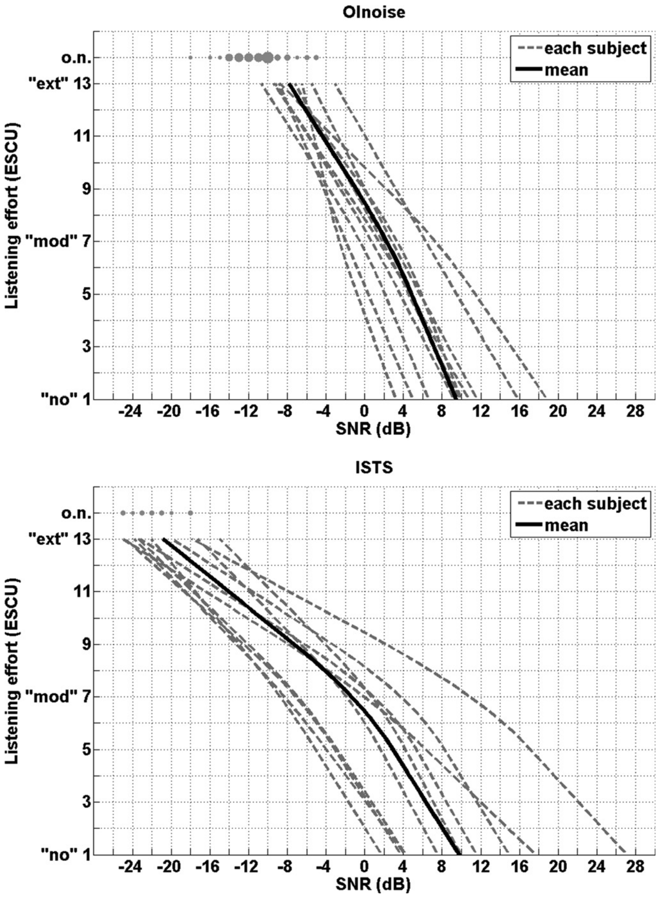

FIG. 4. Individual estimated two-slope functions for the rated listening effort of each subject (dashed grey lines) and mean estimated two-slope functions (solid black lines) using the maskers Olnoise (left) and ISTS (right) and the adaptive procedure. The marker size for the rating category "only noise" (o.n.) visualizes the frequency of use for this response.

fitting the two-slope model function to the data compared to calculating the mean across the data points is that the SNR can be extrapolated for each rating category even if not enough data points are available for robust calculations of the mean in single subject results.

## B. Comparison of initial SNRs

Figure 6 shows the overall estimated two-slope functions for Olnoise and ISTS when using different initial SNRs in the adaptive procedure of experiment 1. For Olnoise, the rating of perceived listening effort was almost identical for the different initial SNRs, especially for the rating category "extreme effort." With decreasing effort, the distance between the regression lines increased slightly. Even in the area of the rating category "no effort," however, the difference between the two-slope functions for the four initial SNRs was maximally 2 dB. The SNR for each of the seven labeled categories calculated from the estimated two-slope functions of each individual subject and each initial SNR was used as an input for the ANOVA for repeated measures. No significant effect of the initial SNR [Greenhouse-Geisser $\varepsilon = 0.540$, $F(1.621, 14.590) = 1.129$, $p = 0.355$] and no interaction between the initial SNR and the rating categories [Greenhouse-Geisser $\varepsilon = 0.149$, $F(2.684, 24.160) = 1.633$, $p = 0.211$] was found.

The overall estimated two-slope functions for ISTS for the different initial SNR were also close to each other. The maximum difference was approximately 2 dB for the rating categories "extreme effort" and "no effort." No significant differences between the SNR values for each of the seven named rating categories for the initial SNRs were found [ANOVA for repeated measurements: $F(2.759, 24.828) = 2.343$, $p = 0.102$] and also no interaction between initial SNR and rating category was found [$F(18, 162) = 1.199$, $p = 0.268$].

## C. Comparison of procedures

In Fig. 7, the estimated two-slope functions for both measurement procedures, adaptive and static, respectively, are shown for each subject and for both maskers. For this comparison, the run of the adaptive procedure with the same initial SNR as the randomly selected initial SNR in the static procedure was used. Irrespective of the procedure used, the quality measure $R_p^2$ was above 0.734 in all cases.

In most of the cases, the comparison of the respective estimated two-slope functions for both procedures revealed similar results. Input for an ANOVA for repeated measures were the SNR values for each category calculated from the estimated two-slope function of each measurement procedure (adaptive or static). The ANOVA supported the similarity between the procedures [$F(1, 9) = 0.962$, $p = 0.352$ for ISTS and $F(1, 9) = 1.402$, $p = 0.267$ for Olnoise]. Nevertheless, for three subjects (S1, S2, S5), the predetermined SNR range in the static procedure did not cover the entire scale from "no effort" (1 ESCU) to "extreme effort" (13 ESCU). Hence, several categories were not used as responses by these subjects. The most extreme case, subject S2, did not use the rating categories from 1 ESCU to 6 ESCU during the static procedure. The adaptive procedure avoided this shortcoming by adjusting the SNR range during the run and resulted in responses for categories not covered by the static procedure that used predefined SNR ranges.

## D. Comparison of maskers

The results of the adaptive procedure used for four maskers in experiment 2 are shown as overall estimated two-slope functions in Fig. 8. The perceived listening effort depended on the masker. Within fluctuating and stationary masker types, listening effort was rated for both maskers in each type as similarly effortful and the estimated two-slope functions were approximately parallel. In contrast, in comparison to those of the stationary maskers, the estimated two-slope functions of the fluctuating maskers were shifted to lower SNR values. Especially in the case of higher listening effort, the same SNR value was perceived as less effortful in fluctuating than in stationary maskers. The difference between the estimated two-slope functions amounted to up to about 10 dB for the rating category "extreme effort." The difference between the estimated two-slope functions decreased with decreasing listening effort and resulted in approx. 3 dB for the category "no effort." The differences between the SNRs of the rating categories for the four maskers were statistically significant [ANOVA for repeated measurements: Greenhouse-Geisser $\varepsilon = 0.613$, $F(1.838,$

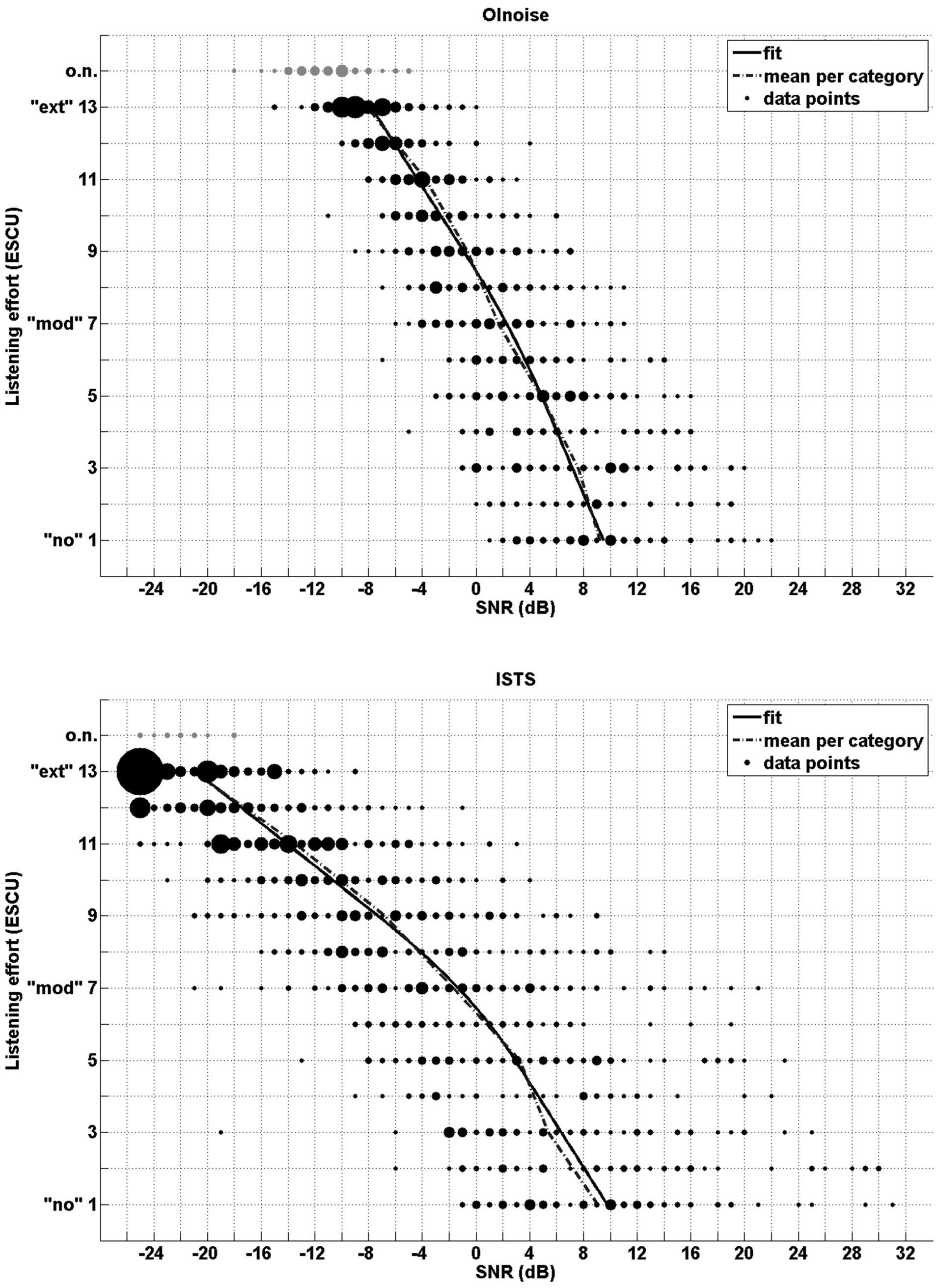

FIG. 5. Overall estimated two-slope function (solid line) and the mean SNR values in each rating category (dashed line) using the maskers Olnoise (top) and ISTS (bottom). The marker size visualizes the frequency of use for each response.

23.890) = 28.882, $p < 0.001$] and also an interaction between maskers and rating category was found [Greenhouse-Geisser $\varepsilon = 0.132$, $F(2.380, 30.934) = 38.861$, $p < 0.001$]. *Post hoc t*-tests, taking the Bonferroni correction into account, showed that the results of the fluctuating maskers were significantly different to the results of the stationary maskers ($p < 0.001$). Within the respective sets, however, no statistically significant differences were ascertainable.

### E. Intra-individual and inter-individual standard deviations

For the calculation of the intra- and inter-individual standard deviations for each masker, the SNR values for the rating categories "no effort," "very little effort," "little effort," "moderate effort," "considerable effort," "very much effort," and "extreme effort" taken from the fitted curves were used. The intra-individual standard deviation was calculated from the three measurement sessions for each subject. Results for one rating category and one masker condition were determined by averaging the intra-individual standard deviations of all subjects (see Table IV, left columns). In comparison, the inter-individual standard deviation was calculated for each session and averaged over all sessions (see Table IV, right columns). With increasing effort, the intra-individual standard deviations and the inter-individual standard deviations decreased until the minimum was achieved at the rating categories "considerable effort" to "very much effort." Taking all maskers together, the largest differences (of about 4 dB) between the two types of standard deviations were observed independent of the masker type for the rating category "no effort." The smallest differences of around 0.8 dB and 2.5–3.0 dB were observed for the stationary and the fluctuating maskers at the rating category "extreme effort," respectively.

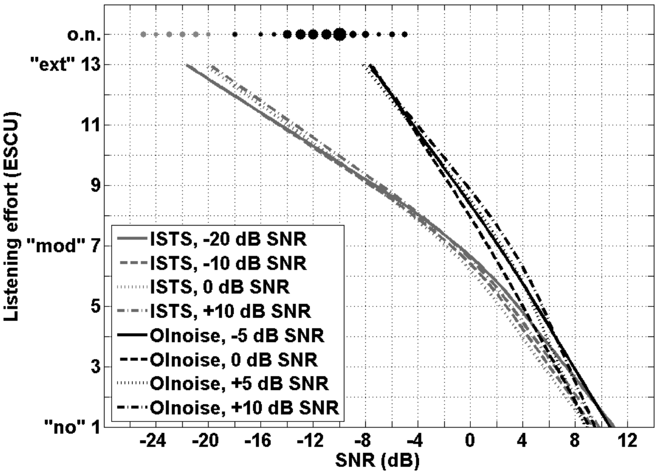

FIG. 6. Overall estimated two-slope functions for different initial SNRs in Olnoise (black lines) and in ISTS (grey lines). The different line styles represent different initial SNRs. For ISTS the initial SNRs were −20 dB, −10 dB, 0 dB, and 10 dB. Initial SNRs for Olnoise were −5 dB, 0 dB, +5 dB, and +10 dB.

### F. Test-retest reliability

Figure 9 shows overall estimated two-slope functions for all four maskers and all three sessions. For all maskers, the estimated two-slope functions for all sessions were very similar, but were slightly shifted for the first session relative to session two and three for the maskers Olnoise and IFFM. The results for the three sessions for the masker IFFM showed the largest difference, with about 3.4 dB at the rating category "moderate effort" (7 ESCU). The highest similarity was observed for the cafeteria masker, with a difference between the regression lines of less than 1.3 dB.

An ANOVA for repeated measures supported the similarity between the SNRs of the rating categories for Olnoise, Cafeteria, and Icra5–250 [Greenhouse-Geisser $\varepsilon = 0.716$, $F(1.432,\ 18.619) = 1.905$, $p = 0.183$ for Olnoise, $F(2,\ 26) = 0.010$, $p = 0.990$ for Cafeteria, and $F(2,\ 26) = 0.179$, $p = 0.837$ for Icra5–250] and revealed a statistically significant difference for IFFM [$F(2,\ 26) = 4.673$, $p = 0.018$]. A *post hoc* paired comparison *t*-test with the adjusted significance level using the Bonferroni correction supported a statistically significant differences between the first and second session for IFFM ($p = 0.041$).

The intra-class correlation coefficient (see Table V) supported the similarity between the estimated two-slope functions for all three sessions ($p < 0.002$ for all conditions). Most of the coefficients were above 0.84, with three exceptions for the rating category "very much effort" (IFFM and Icra5-250) and "extreme effort" (Icra5–250). The mean intra-class correlation coefficient for all maskers was about 0.87. A comparison of the mean intra-class correlation coefficient of each masker revealed the highest value (0.923) for the masker Olnoise and the lowest value (0.831) for the Icra5-250 masker.

## IV. DISCUSSION

### A. Individual fits

For the aggregation of the subjectively perceived listening effort for each individual listener in each masker, two-slope functions were fitted to the data points using the BX model. The high $R_p^2$ values as representative measures for the goodness of fit for each subject showed that the estimated two-slope functions using the BX fitting method resulted in a satisfactory approximation. Only for one subject were $R_p^2$ values below 0.8 observed for both maskers. Individual observation of this subject's ratings as a function of the SNRs did not reveal any systematic deviations, but rather a broader scatter of the ratings in comparison to other subjects; this had a negative impact on the fitting of the two-slope function. Therefore, future implementations of this adaptive method might automatically delete outliers, to improve the estimations of the two-slope functions. In general, however, a two-slope function was appropriate to represent the data. The form of the function leads to consider that effort is not a linearly scaled construct at least for part of the listeners and in some masker conditions. Effort can grow more rapidly or more slowly at different parts of the effort range. Alternatively, the perceptual labeling of effort ratings might not be linear instead of the effort itself. Those two possibilities cannot be distinguished based on the data in this study.

The present contribution only shows data for NH subjects, and it cannot be concluded from these results that data from HI subjects can also be approximated by a two-slope function. The similar fitting function for NH and HI subjects in the ACALOS procedure (Brand and Hohmann, 2002) may, however, indicate the possible applicability of the two-slope function for HI listeners in the ACALES procedure.

### B. Comparison of initial SNRs

The goal of the first phase of the adaptive procedure was to find the SNRs of the boundaries that correspond to the ratings "no effort" and "extreme effort." To explore whether an initial SNR value for the first presentation near one of the boundaries might influence the outcomes, different initial SNRs were investigated. A comparison for each masker showed that the initial SNR had no significant effect on the subjective listening effort ratings. This indicates that the initial SNR is not used as an anchor for medium ratings and it can be chosen without influencing the results. This result might be especially important for HI listeners with different hearing loss and different speech recognition in noise. Nevertheless, for comparison reasons, using the same initial SNR for the evaluation of subjective behavior or hearing aid algorithms is recommended.

### C. Comparison of procedures

The comparison of the respective estimated two-slope functions for both procedures revealed no statistically significant differences between the static procedure with constant stimuli and the adaptive procedure. Nevertheless, due to the large inter-individual differences in subjectively perceived listening effort, the predetermined SNR range in the static procedure did not cover the entire range of possible listening effort ratings from "no effort" to "extreme effort" for three subjects. Consequently, these subjects did not use the categories between "no effort and "very little effort" or "moderate effort." This response pattern deviated from the expectation that subjects tend to use the entire range of the

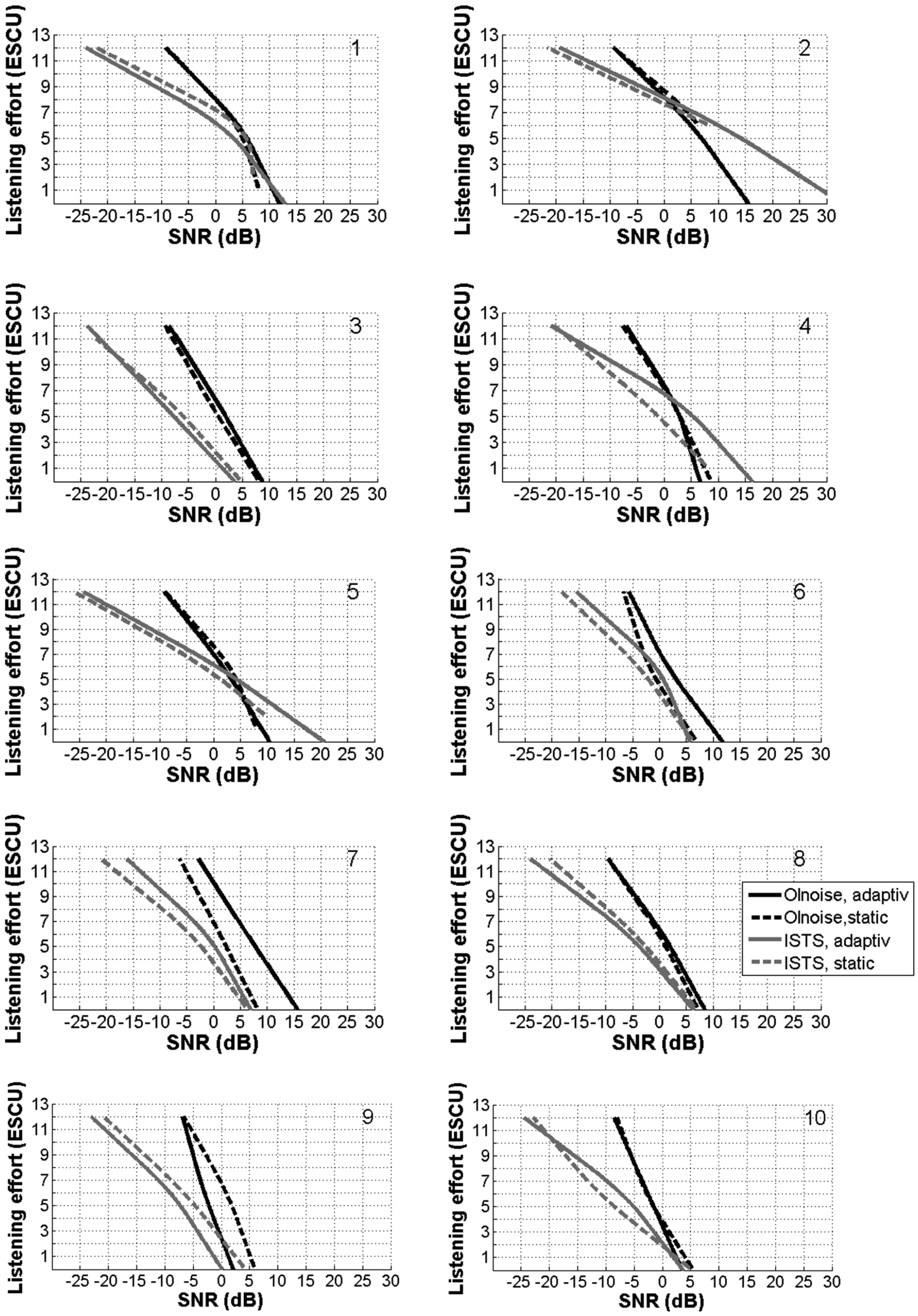

FIG. 7. Comparison of the adaptive (solid lines) and static (dashed lines) procedure in Olnoise (black) and ISTS (grey) for each subject. The rating category "no effort" corresponds to 1 ESCU, "moderate effort" to 7 ESCU and "extreme effort" to 13 ESCU. In this figure, ratings for "only noise" are not presented, to improve visibility of the data.

response scale for the presented range of stimuli, resulting in a possible bias if both ranges do not match (e.g., see Zielinski, 2016). Both effects, coverage of only a part of the response scale and response bias, are shortcomings of the static procedure, which were resolved by the adaptive procedure using individually determined SNR ranges.

An advantage of the new adaptive procedure is that no separate pretests are necessary to determine the required range of SNRs. van Schoonhoven *et al.* (2016) used the static version of the categorical ratings of listening effort and other methods, such as localization and speech intelligibility, to evaluate the benefit of binaural amplification with different noise setups for NH and two groups of HI listeners. For each group of subjects, they used a different static SNR range. However, in their study, the listening effort ratings showed the largest effect sizes for a binaural benefit for subjects with mild hearing losses. Since van Schoonhoven *et al.* (2016) demonstrated the applicability of subjective listening effort ratings for NH and HI subjects in the evaluation of hearing aid algorithms, a similar applicability is expected for the ACALES method while avoiding preselection of the SNR range.

## D. Comparison of maskers

The adaptive procedure for listening effort showed that fluctuating and stationary maskers provoke significantly different ratings, especially in lower SNR ranges, resulting in different slopes for these masker types. This was in agreement with speech intelligibility measurements, which typically result in lower slopes for fluctuating and higher slopes

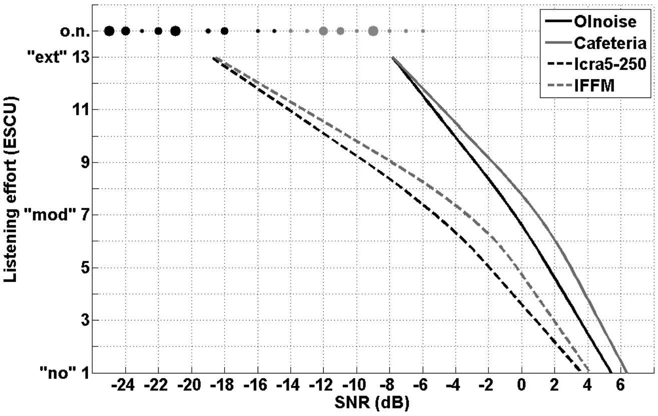

FIG. 8. Comparison of rated listening effort for Cafeteria (solid grey line), Icra5-250 (dashed black line), IFFM (dashed grey line), and Olnoise (solid black line). The black dots represent the use of rating category "only noise" (o.n.) for the fluctuating maskers (Icra5-250 and IFFM) and the grey dots for the stationary maskers (Olnoise and Cafeteria).

for stationary maskers (Wagener and Brand, 2005). The two maskers of each masker type (fluctuating or stationary) showed no significant differences in terms of the rated listening effort, although the masker's content, especially for the fluctuating type, was quite different.

Differences in subjective listening effort between speech as a masker and continuous speech-shaped noise were also reported by Hällgren *et al.* (2005). However, they reported higher effort ratings for the fluctuating speech as a masker. A possible explanation might be that they presented the stimuli at a rather positive SNR of +10 dB. In this study, subjective listening effort is lower in the fluctuating masker IFFM compared to the stationary masker Olnoise. However, the difference seems to be larger for lower than for higher SNRs. A similar result was found by Devocht *et al.* (2016) for bimodal CI users. They presented sentences of the Dutch matrix test in fluctuating and stationary noise. In agreement with our study, they found lower ratings for the fluctuating masker at low SNRs and similar ratings for high SNRs and suggested that the relationship between intelligibility and listening effort is quite different for stationary and fluctuating noise.

### E. Intra-individual and inter-individual standard deviations

The inter-individual standard deviation of 7.2 dB for the NH listeners tested in this study was much larger than the intra-individual standard deviation of 2.9 dB. As the intra-individual is nearly three times smaller than the inter-individual standard deviation we conclude that individual differences in perception can be resolved by the ACALES procedure.

For a categorical loudness scaling procedure with a numerical scale from 1 to 8, Robinson and Gatehouse (1996) observed an intra-listener standard deviation for NH subjects between 3.5 and 7.3 dB, depending on the stimulus level. They regarded these deviations as good enough to recommend the application of the procedure. Since the intra-individual standard deviation of the ACALES procedure was even smaller than found for categorical loudness scaling, their recommendation should also apply here.

Johnson *et al.* (2015) demonstrated that self-reported measures of listening effort were a reliable measure to distinguish between SNRs. They used the same categorical scale as in this study for subjective listening effort ratings during a speech recognition test and compared it with a word recall task. As the subjective method was more sensitive to changes in SNR (in steps of 2 dB), they provided a rationale for preferring the self-report measure of listening effort over the word recall measure.

The cause of the large inter-individual differences is unclear. Even when using the same instructions, the inter-individual differences in SNR values, particularly for the rating category "no effort," revealed deviations in the criterion for an effortless situation. This criterion was individually established, since no anchor was provided with the instructions. At the other end of the scale, the category beyond "extreme effort," which was labeled as "only noise" and was meant for situations when no speech cues were perceived, might be regarded as an anchor. The listeners were informed that a listening situation could be regarded as "effortless" when they were able to follow the speaker for a longer time period without effort, but comparisons such as, e.g., "as effortless as…" were not presented. The inter-individual differences might be reduced by offering examples for different categories, e.g., for "no effort" and "extreme effort," but these examples might provoke bias and were therefore avoided.

### F. Test-retest reliability

The mean intra-class correlation coefficient for all noises was above 0.9. Overall, the correlation was higher for lower effort ratings, which correspond to higher SNRs. When

TABLE IV. Results of the intra- and the inter-individual standard deviation for the rating category "no effort," "very little effort," "little effort," "moderate effort," "considerable effort," "very much effort," and "extreme effort" in four maskers.

| | Intra-individual standard deviation in dB (mean) | | | | | Inter-individual standard deviation in dB (mean) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Olnoise | Cafeteria | IFFM | Icra5-250 | Mean | Olnoise | Cafeteria | IFFM | Icra5-250 | Mean |
| No effort | 1.9 | 2.5 | 3.8 | 2.5 | 2.7 | 5.9 | 5.8 | 7.8 | 6.5 | 6.5 |
| Very little effort | 1.5 | 2.0 | 3.3 | 2.2 | 2.2 | 5.1 | 5.0 | 7.4 | 6.2 | 6.0 |
| Little effort | 1.2 | 1.6 | 2.8 | 2.1 | 1.9 | 4.5 | 4.4 | 7.2 | 6.2 | 5.6 |
| Moderate effort | 1.2 | 1.3 | 2.5 | 2.2 | 1.8 | 3.8 | 3.7 | 6.8 | 6.0 | 5.1 |
| Considerable effort | 1.2 | 1.2 | 2.4 | 2.3 | 1.7 | 2.9 | 2.7 | 5.9 | 5.5 | 4.3 |
| Very much effort | 1.0 | 1.1 | 2.5 | 2.3 | 1.7 | 2.0 | 2.0 | 5.3 | 5.2 | 3.6 |
| Extreme effort | 1.0 | 1.3 | 2.8 | 2.4 | 1.9 | 1.9 | 2.0 | 5.3 | 5.4 | 3.7 |

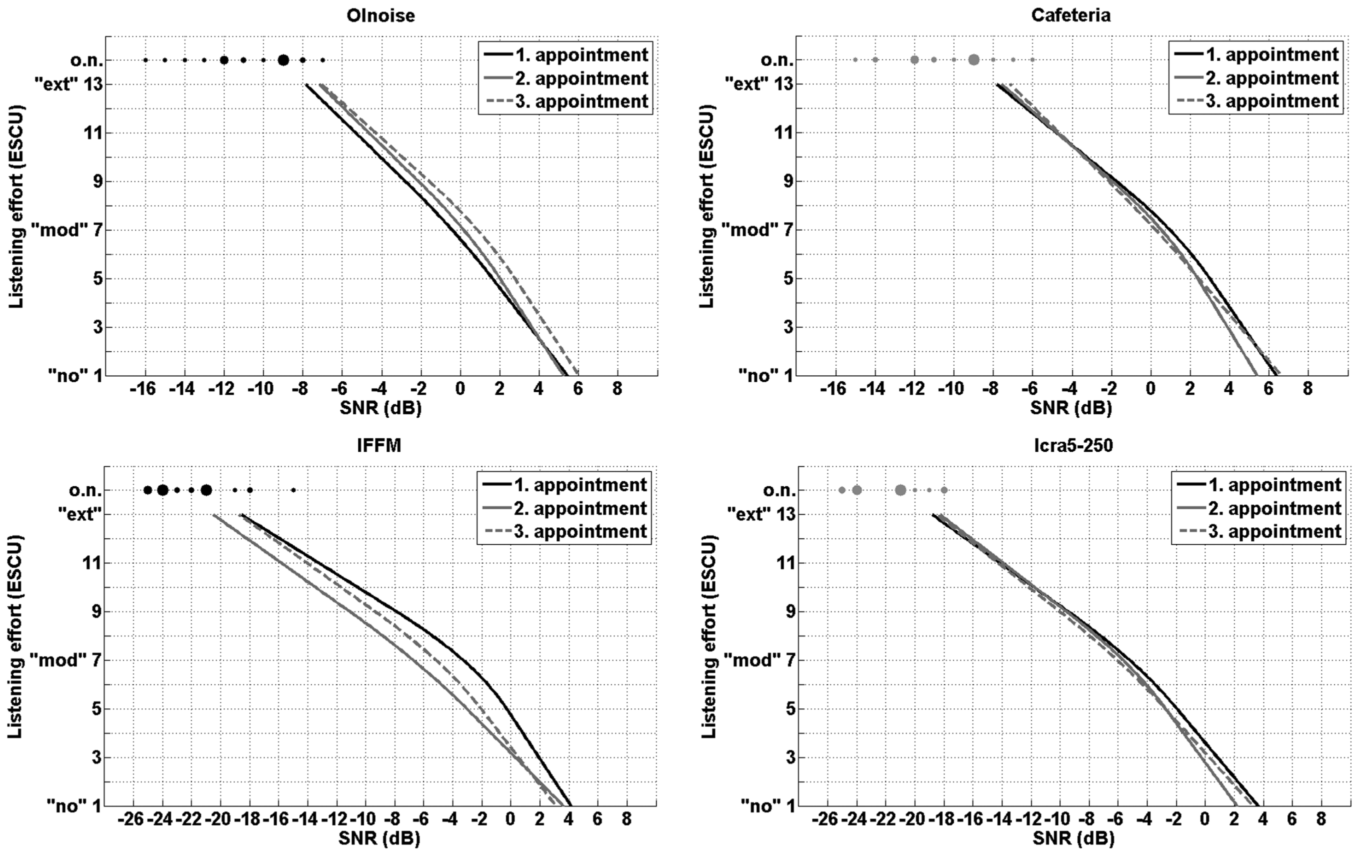

FIG. 9. Comparison of the results measured on the first (black line), second (grey line), and third (dashed grey line) session for Olnoise, Cafeteria, IFFM and Icra5-250 maskers.

using a categorical loudness scaling procedure, Robinson and Gatehouse (1996) also observed an improvement in intra-listener standard deviation with increasing level from 7.3 to 3.5 dB. A lower test-retest reliability for lower SNRs seems to contradict a possible anchor-effect at "extreme effort" or "only noise," as discussed in the previous paragraph. On the other hand, the OLSA is known to be affected by learning effects (Schlueter *et al.*, 2016). An increased familiarization with the speech material might impact ratings in the area of the threshold for speech recognition more than in the area of speech recognition scores of 100%.

Luts *et al.* (2010) used the static version of the listening effort rating procedure. They evaluated different single-channel noise reduction algorithms by means of speech intelligibility tests, preference ratings, and listening effort ratings. In the listening effort and preference ratings, but not in the intelligibility tests, a subjective benefit of single channel noise reduction was found. All measurements were carried out during two visits and the test-retest reliability was calculated for distinct SNRs and ranged from 0.6 to 1.2 scale units. As they transformed the ratings on the same categorical scale to values from 0 to 6, instead of from 1 to 13 as in our study, this corresponds to 1.2 to 2.4 ESCUs. In the present analysis, the SNR axis was used to calculate test-retest differences in dB SNR for the rating category "no effort" instead of calculating the test-retest reliability in ESCUs. The differences ranged from 1.7 to 3.7 dB. However, when using the slope of the mean functions for the different maskers, this range in dB corresponds to a test-retest variability of 1.0 to 3.3 ESCUs, which is in reasonable agreement with Luts' data.

## V. CONCLUSIONS

A new method for measuring self-reported listening effort was presented: the Adaptive Categorical Listening

TABLE V. Intra-class correlation coefficients calculated with the "two way mixed" model and the "absolute agreement" type for average measures.

| | Intra-class correlation coefficient for rating category | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | No effort | Very little effort | Little effort | Moderate effort | Considerable effort | Very much effort | Extreme effort | Mean |
| Olnoise | 0.929 | 0.953 | 0.964 | 0.952 | 0.926 | 0.876 | 0.858 | 0.923 |
| Cafeteria | 0.853 | 0.887 | 0.918 | 0.930 | 0.916 | 0.840 | 0.754 | 0.871 |
| IFFM | 0.869 | 0.902 | 0.920 | 0.916 | 0.880 | 0.808 | 0.743 | 0.863 |
| Icra5-250 | 0.896 | 0.908 | 0.905 | 0.879 | 0.820 | 0.732 | 0.676 | 0.831 |

Effort Scaling (ACALES). The new method was tested with two groups of NH subjects in four different background maskers. The following conclusions can be made:

- The ACALES procedure is easy and fast. It was accomplishable by all subjects and no separate pretests were required to determine a valid SNR range.
- The procedure captures individual differences in subjective listening effort. Hence, more data in the relevant range of SNRs can be collected and a detailed overview of the individually perceived range of listening effort is obtained.
- The results are independent of the initial SNR, facilitating the applicability to different groups of subjects.
- The procedure is able to resolve differences between stationary and fluctuating maskers, revealing that the stationary maskers are subjectively perceived as more effortful.

## ACKNOWLEDGMENTS

We thank all listeners for their participation in the study and Marit Nienaber for contributing to data collection. English language support was provided by www.stels-ol.de. Supported by the governmental funding initiative "Niedersächsisches Vorab" of the Lower Saxony Ministry for Science and Culture, research focus "Hören im Alltag Oldenburg (HALLO)" as well as by the European Regional Development Fund [ERDF-Project "Innovation network for integrated, binaural hearing system technology (VIBHear)"] and funds from the Federal State of Lower Saxony.

Brand, T., and Hohmann, V. (**2002**). "An adaptive procedure for categorical loudness scaling," J. Acoust. Soc. Am. **112**, 1597–1604.

Devocht, E. M., Janssen, A. M., Chalupper, J., Stokroos, R. J., and George, E. L. (**2016**). "Monaural beamforming in bimodal cochlear implant users: Effect of (a)symmetric directivity and noise type," PLoS One. **11**(8), e0160829.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "Assessment: Ruidos ICRA: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos" ("ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument"), Audiology **40**, 148–157.

Gatehouse, S., and Noble, W. (**2004**). "The speech, spatial and qualities of hearing scale (SSQ)," Int. J. Audiol. **43**, 85–99.

Hällgren, M., Larsby, B., Lyxell, B., and Arlinger, S. (**2005**). "Speech understanding in quiet and noise, with and without hearing aids," Int. J. Audiol. **44**, 574–583.

Hart, S. G., and Staveland, L. E. (**1988**). "Development of NASA-TLX (task load index): Results of empirical and theoretical research," Adv. Psychol. **52**, 139–183.

Holube, I. (**2011**). "Speech intelligibility in fluctuating maskers," in *Speech Perception and Auditory Disorders*, edited by T. Dau, M. L. Jepsen, T. Poulsen, and J. C. Dalsgaard (The Danavox Jubilee Foundation, Ballerup, Denmark), pp. 57–64.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an international speech test signal (ISTS)," Int. J. Audiol. **49**, 891–903.

Holube, I., Haeder, K., Imbery, C., and Weber, R. (**2016**). "Subjective listening effort and electrodermal activity in listening situations with reverberation and noise," Trends Hear **20**, 1–15.

Johnson, J., Xu, J., Cox, R., and Pendergraft, P. (**2015**). "A comparison of two methods for measuring listening effort as part of an audiologic test battery," Am. J. Audiol. **24**, 419–431.

Klink, K. B., Schulte, M., and Meis, M. (**2012a**). "Measuring listening effort in the field of audiology – a literature review of methods, part 1," Z. Audiol. **51**, 60–67.

Klink, K. B., Schulte, M., and Meis, M. (**2012b**). "Measuring listening effort in the field of audiology – a literature review of methods, part 2," Z. Audiol. **51**, 96–106.

Larsby, B., Hällgren, M., Lyxell, B., and Arlinger, S. (**2005**). "Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects," Int. J. Audiol. **44**, 131–143.

Lemke, U., and Besser, J. (**2016**). "Cognitive load and listening effort: Concepts and age-related considerations," Ear Hear **37**, 77S–84S.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, M., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., and Spriet, A. (**2010**). "Multicenter evaluation of signal enhancement algorithms for hearing aids," J. Acoust. Soc. Am. **127**, 1491–1505.

Mackersie, C. L., and Calderon-Moultrie, N. (**2016**). "Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance," Ear Hear **37**, 118S–125S.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., and Amitay, S. (**2014**). "Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper,' " Int. J. Audiol. **53**, 433–445.

McShefferty, D., Whitmer, W. M., and Akeroyd, M. A. (**2015**). "The just-noticeable difference in speech-to-noise ratio," Trends Hear. **19**, 2331216515572316.

Montgomery, H. (**1975**). "Direct estimation: Effect of methodological factors on scale type," Scand. J. Psychol. **16**, 19–29.

Oetting, D., Brand, T., and Ewert, S. D. (**2014**). "Optimized loudness-function estimation for categorical loudness scaling data," Hear. Res. **316**, 16–27.

Parducci, A., and Perret, L. F. (**1971**). "Category rating scales: Effect of relative spacing and frequency of stimulus values," J. Exp. Psychol. Monogr. **89**, 427–452.

Pichora-Fuller, K. M., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., and Wingfield, A. (**2016**). "Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL)," Ear Hear. **37**, 5S–27S.

Poulton, E. C. (**1989**). *Bias in Quantifying Judgements* (Lawrence Erlbaum, Hillscale, NJ), 275 p.

Rennies, J., Schepker, H., Holube, I., and Kollmeier, B. (**2014**). "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," J. Acoust. Soc. Am. **136**, 2642–2653.

Robinson, K., and Gatehouse, S. (**1996**). "Test-retest reliability of loudness scaling," Ear Hear. **17**, 120–123.

Sato, H., Bradley, J. S., and Morimoto, M. (**2005**). "Using listening difficulty ratings of conditions for speech communication in rooms," J. Acoust. Soc. Am. **117**, 1157–1167.

Schepker, H., Haeder, K., Rennies, J., and Holube, I. (**2016**). "Listening effort and speech intelligibility in reverberation and noise for hearing-impaired listeners," Int. J. Audiol. **55**, 738–747.

Schlueter, A., Lemke, U., Kollmeier, B., and Holube, I. (**2016**). "Normal and time-compressed speech: How does learning affect speech recognition thresholds in noise?," Trends Hear. **20**, 1–13.

Schulte, M., Vormann, M., Wagener, K. C., Buchler, M., Dillier, N., Dreschler, W., and Wouters, J. (**2009**). "Listening effort scaling and preference rating for hearing aid evaluation," in *HearCom Workshop on Hearing Screening and Technology*, Brussels, Belgium, http://hearcom.eu/lenya/hearcom/authoring/about/DisseminationandExploitation/Workshop/S2B-3_Michael-Schulte_Hearing-Aid-Scaling-Rating.pdf (Last viewed January 20, 2015).

van Schoonhoven, J., Schulte, M., Boymans, M., Wagener, K. C., Dreschler, W. A., and Kollmeier, B. (**2016**). "Selecting appropriate tests to assess the benefits of bilateral amplification with hearing aids," Trends Hear. **20**, 1–16.

Wagener, K., Brand, T., and Kollmeier, B. (**1999a**). "Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test," Z. Audiol. **38**, 44–56.

Wagener, K., Brand, T., and Kollmeier, B. (**1999b**). "Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test," Z. Audiol. **38**, 86–95.

Wagener, K., Kühnel, V., and Kollmeier, B. (**1999c**). "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test," Z. Audiol. **38**, 4–15.

Wagener, K. C., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters," Int. J. Audiol. **44**, 144–156.

Wagener, K. C., Brand, T., and Kollmeier, B. (**2006**). "The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners," Int. J. Audiol. **45**, 26–33.

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

Zielinski, S. (**2016**). "On some biases encountered in modern audio quality listening tests (part 2): Selected graphical examples and discussion," J. Audio Eng. Soc. **64**, 55–74.