Original Article

Trends in Hearing
Volume 22: 1–13

DOI: 10.1177/2331216518802702
journals.sagepub.com/home/tia
SAGE

The Self-Assessed Békesy Procedure: Validation of a Method to Measure Intelligibility of Connected Discourse

Lien Decruy[1], Neetha Das[1,2], Eline Verschueren[1], and Tom Francart[1]

Abstract

In clinical practice and research, speech intelligibility is generally measured by instructing the participant to recall sentences. Although this is a reliable and highly repeatable measure, it cannot be used to measure intelligibility of connected discourse. Therefore, we developed a new method, the self-assessed Békesy procedure, which is an adaptive procedure that uses intelligibility ratings to converge to a person's speech reception threshold. In this study, we describe the new procedure and the validation in young, normal-hearing listeners. First, we compared the results on the self-assessed Békesy procedure to a recall procedure for standardized sentences. Next, we evaluated the inter- and intrasubject variability of our procedure. Furthermore, we compared the thresholds for sentences in three masker types between the self-assessed Békesy and a recall procedure to verify if these procedures resulted in similar conclusions. Finally, we compared the thresholds for two types of sentences and commercial recordings of stories. In general, the self-assessed Békesy procedure is shown to be a valid and reliable procedure as similar thresholds (difference < 1 dB) and test–retest reliability (< 1.5 dB) were observed compared with standard speech audiometry tests. In addition, the time efficiency and similar differences between maskers to a recall procedure support the potential of this procedure to be implemented in research. Finally, significant differences between the thresholds of sentences and connected discourse materials were found, indicating the importance of controlling for differences in intelligibility when presenting these materials at the same signal-to-noise ratios or when comparing studies.

Keywords

connected discourse, speech intelligibility, speech audiometry, adaptive procedure, continuous speech

Date received: 17 January 2018; revised: 27 August 2018; accepted: 29 August 2018

Introduction

Speech intelligibility is usually measured using standardized speech materials where participants are asked to recall words or sentences they heard. The responses of the participants are then scored per word or per sentence depending on the type of speech material. For speech audiometry, two procedures are often used: the constant and adaptive procedure. With the constant procedure, a list of words or sentences is presented at a particular signal-to-noise ratio (SNR) whereas the adaptive procedure converges to a certain speech intelligibility level by changing the SNR based on the participant's response (Levitt, 1971). The convergence point for the adaptive procedure is often the 50% speech intelligibility level, also called the speech reception threshold (SRT). Although speech audiometry is a reliable and highly repeatable measure, it is not applicable to considerably longer sentences, especially connected discourse, as these cannot be easily recalled and scored in the same way as short sentences. Furthermore, connected discourse materials are usually not standardized. Despite this, commercial or lab recordings of stories have been used in studies to evaluate a person's speech intelligibility during a behavioral or electrophysiological test (Ding

[1]ExpORL, Department of Neurosciences, KU Leuven—University of Leuven, Belgium
[2]ESAT, Department of Electrical Engineering, KU Leuven—University of Leuven, Belgium

Corresponding author:
Tom Francart, Department of Neurosciences, ExpORL, KU Leuven – University of Leuven, Herestraat 49 Bus 721, B-3000 Leuven, Belgium.
Email: tom.francart@kuleuven.be

& Simon, 2013; Falconer & Davis, 1947; MacPherson & Akeroyd, 2014; Petersen, Wöstmann, Obleser, & Lunner, 2017). As there are large differences regarding content and acoustic properties between these stories, it is difficult to compare the results between these studies. Therefore, other methods are needed to investigate the intelligibility of connected discourse.

Several procedures have been used to measure intelligibility of connected discourse. One method is the content-related question-and-answer procedure. Here, the participant is asked comprehension questions during or after listening to a passage of speech. By counting the number of correct answers, speech intelligibility is measured (e.g., Petersen et al., 2017). Best, Keidser, Buchholz, and Freeston (2016) showed that the SRTs of sentence tests were highly correlated with those on their content-related question-and-answer test. Although this method seems promising, it is time consuming since a constant procedure is used and the results highly rely on the listener's postperceptual abilities (MacPherson & Akeroyd, 2014), the content of the speech material and how the content-related questions are selected.

A second method is the speech Békesy procedure where the participants have to adjust the level of the masker or target stimulus until a certain level of speech intelligibility is reached. Studies have shown that this adaptive, time efficient procedure has a good test–retest reliability and keeps the participants motivated and alert as they are actively self-adjusting the level (Falconer & Davis, 1947; Speaks, Trine, Crain, & Niccum, 1994; review of Kei, Smyth, Murdoch, & McPherson, 1999). Falconer and Davis (1947) developed a test called the threshold of intelligibility for connected discourse (TICD) and instructed the participants to adjust the level of a newscast, until some of the words dropped out. They found a difference of only 0.8 dB between the norm SRT of a word test (22.5 dB) and the TICD (23.2 dB). The average test–retest difference was smaller for the TICD (1.8 dB) compared with the word test (2.1 dB).

In clinical practice as well as for research purposes, the self-assessed procedure is often used (Anderson-Gosselin & Gagne, 2010; Ding & Simon, 2013; Gatehouse & Noble, 2004; Zekveld, Kramer, & Festen, 2010). In this procedure, participants are instructed to rate the intelligibility of sentences or connected discourse by giving a percentage or using a scale during a constant procedure. Walker and Byrne (1985) reported a good test–retest reliability of rating by normal-hearing participants (median test–retest difference: 1.2 dB). Despite this, it should be noted that the underlying principle of a person's self-adjusting (speech Békesy procedure) or rating strategy (self-assessed procedure) can bias the results (Falconer & Davis, 1947). For example, a strategy based on *accurately perceiving words* may result in similar outcomes as standard recall sentence tests whereas a strategy based on *understanding the main message* can result in different outcomes or a higher intersubject variability.

Recently, a new method was developed by MacPherson and Akeroyd (2014), the Glasgow monitoring of uninterrupted speech task. In this procedure, the participant is instructed to listen to connected discourse while simultaneously reading a written transcript. The transcript contains deliberate mistakes and the number of detected substitutions can be taken as a measure of speech intelligibility. They found a worse SRT, shallower psychometric function, and higher intersubject variability for continuous connected discourse compared with trial-by-trial recalled sentences. Furthermore, the results for two extra conditions, trial-by-trial connected discourse and concatenated sentences, suggest that the higher difficulty of connected discourse was not related to the duration of the stimuli. However, in order to correctly interpret the difference in dB between sentences and connected discourse, concatenated sentences should be used as a better SRT was obtained when presenting sentences continuously instead of trial-by-trial. This task also has several drawbacks. For example, the authors report that participants need to have an adequate reading ability. Furthermore, the amount of top-down information that can be used in this test is higher compared with methods without written transcripts since it provides the participant extra knowledge. This makes the test rather unrealistic and may underestimate the difficulty associated with the intelligibility of connected discourse.

We developed a new method, aimed to address a number of shortcomings of the methods reviewed earlier, which we term the *self-assessed Békesy procedure*. This is an adaptive procedure in which the level of the masker is adapted based on the participant's rating of speech intelligibility. As a result of this, not only the intelligibility of sentences but also of connected discourse can be evaluated as it converges to the SRT of the participant. Since it is an adaptive and therefore time efficient procedure, we added the term "Békesy" in the name of our procedure. We also included the term "self-assessed" as we use ratings to adapt the SNR. We believe that ratings are more natural compared with self-adjusting the level and also motivate participants to listen to the content of the target speaker instead of relying on the level of the masker. Since we use ratings, we will also refer to this new method as the *rate* procedure. To make a clear distinction between this procedure and a standard adaptive speech audiometry test where persons recall sentences, we will refer to the latter method as the *recall* procedure.

In the remainder of this article, we will describe the self-assessed Békesy procedure and it's validation in a group of young, normal-hearing listeners. An overview

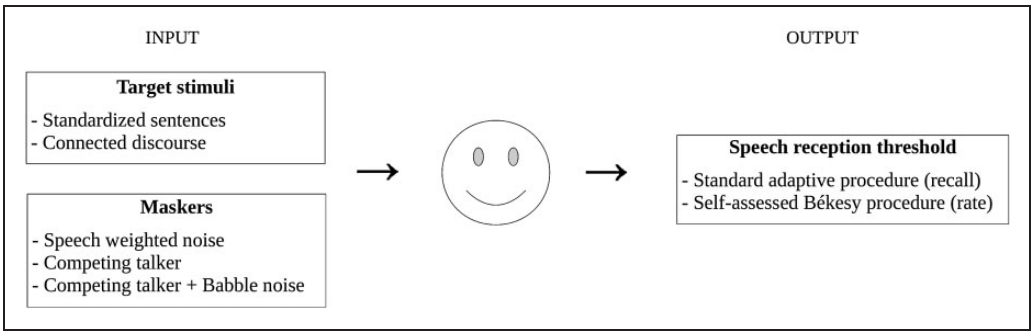

Figure 1. Overview of the study design.

of the study design is given in Figure 1. Using these material and methods, we investigated the following:

- We compared the outcomes of the self-assessed Békesy procedure (i.e., rate procedure) with those of a gold standard speech audiometry test (i.e., recall procedure) to assess the validity of our new procedure. The same standardized sentences were used in both procedures.
- We calculated the inter- and intrasubject variability. To evaluate the test–retest reliability, the rate procedure for sentences and connected discourse was conducted at least two times.
- We included three different maskers in our study to verify if our rate procedure detects similar differences between maskers compared with the recall procedure. This is important since studies have shown that demanding, informational maskers result in more difficulties in older compared with younger normal-hearing participants (Goossens, Vercammen, Wouters, & van Wieringen, 2017). Such cognitive factors can in turn differentially affect the recall versus rate measure. In addition, this allowed us to investigate the effect of masker type on intelligibility of connected discourse, which has not been done before.
- We examined the difference between the rate SRTs of concatenated unrelated sentences (continuous speech) versus connected discourse to evaluate whether it is important to control for differences between sentences and connected discourse when presenting them at the same SNRs.
- We compared the test duration of the recall and rate procedures to formulate an advice about the implementation of this test in an experimental and clinical setting.

Material and Methods

Participants

Fourteen participants aged between 18 and 26 years (12 women and 2 men, median age: 20) participated in our study and had Dutch (Flemish) as their mother tongue. All participants had normal hearing as they had pure tone thresholds better than 25 dB HL at all octave frequencies from 125 Hz up to 8 kHz, in both ears. This study was approved by the Medical Ethics Committee UZ KU Leuven / Research (Reference No. S57102). All participants took part voluntarily and gave their written informed consent. According to a power analysis for pairwise *t*-test comparisons, 14 subjects should be enough to find a large effect (effect size = 0.8) with significance level equal to .05 and power equal to 0.8.

Stimuli

Target stimuli. To compare the recall and rate procedure, we chose two different types of standardized sentences, both uttered by female speakers and normed in a young, normal-hearing population (Luts, Jansen, Dreschler, & Wouters, 2014; van Wieringen & Wouters, 2008). The two types of speech materials were both included in the study because they differ in their resemblance to connected discourse and daily life sentences. The first type of sentences is the Flemish Matrix sentence test (Luts et al., 2014), which contains 13 lists of 20 sentences. Each sentence has the fixed structure *name, verb, numeral, color* and *object* where each element is selected from a closed-set of 10 possibilities, for example, *Jacob ziet drie groene boten* (*Jacob sees three green boats*). The Matrix sentences are grammatically trivial and cannot be completed based on context cues. During a recall procedure, participants are instructed to recall the sentences which allow the percentage of correctly repeated words to be calculated (word scoring).

The second type of sentences is the Leuven intelligibility sentence test (LIST), which consists of 35 lists of 10 sentences each (van Wieringen & Wouters, 2008). This speech material resembles daily life sentences and connected discourse more closely, as it contains context cues, for example, *De bakker bakt brood* (*The baker bakes bread*). For the LIST sentences, sentence scoring based on keywords is used to calculate the percentage of correct recalled sentences. More specifically, a sentence

has a maximum score of 100% if each keyword of the sentence is identified correctly; otherwise, the sentence score is equal to 0%. Errors of non-keywords, such the article *the,* are not taken into account. Since the LIST sentences were originally developed to test hearing aid and cochlear implant users, they are spoken relatively slowly. Therefore, we speeded up the sentences with a factor of 0.75 using the program Praat (Boersma & Weenink, 2013) to more resemble the rates of the Matrix sentences and connected discourse.

To measure the intelligibility of connected discourse, we used two nonstandardized, commercial recordings of stories as target stimuli. The first story is called *De Wilde Zwanen* from Hans Christian Andersen, narrated by a female, Flemish talker (Story1). The second story is called *Bianca en Nero* from Béatrice Deru-Renard and narrated by a male, Flemish talker (Story2). Both stories were set to the same root mean square level, and silences were shortened to a maximum duration of 300 ms.

Maskers. In this study, we used three different maskers: speech weighted noise (SWN), a competing talker (CT), and a competing talker in combination with babble noise (CTBabble) for two reasons. First, to verify if the rate procedure is able to detect differences between maskers similar to the recall procedure. Second, to investigate the effect of masker type on intelligibility of connected discourse. As shown in Table 1, all standardized sentences and stories were presented in stationary SWN. SWN has the long-term average spectrum of the corresponding speech materials or stories and therefore results in optimal spectral masking, also called energetic masking. Besides being used as a target, Story2 was also used as a CT to mask Story1. In addition to energetic masking, a CT also results in the activation of top-down cognitive processes such as selective attention (i.e., informational masking) and makes it difficult to separate the target and CT (for review, see Kidd & Colburn, 2017). Despite this, a CT has temporal gaps and can consequently enable the participant to achieve better SRTs compared with the SWN condition (Francart, van Wieringen, & Wouters, 2011). Lastly, the story *Milan* (Story3), written and narrated by Stijn Vranken, in combination with babble noise (CTBabble) was used to create a condition which is similar to a realistic, challenging communication scenario such as a cafeteria.

Table 1. Overview of the Different Conditions.

	Recall			Rate		
	SWN	CTBabble	CT	SWN	CTBabble	CT
Matrix	X	X	X	X	X	X
LIST	X	X	X	X	X	X
Story1				X		X
Story2				X	X	

Note. SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise; LIST = Leuven intelligibility sentence test.

Apparatus and Presentation of Stimuli

The participant was seated in a triple-walled soundproof booth in front of a computer running the software platform APEX (Francart, van Wieringen, & Wouters, 2008). All stimuli were presented through ER-3 A insertphones while instructions were given on the computer screen. The target stimuli were calibrated at 90 dB SPL (A weighted) with a type 2260 sound level pressure meter, a type 4189 half-inch microphone, and a 2 cc coupler from Brüel & Kjaer. An RME Multiface II soundcard was connected to the computer with a PCMCIA HDSP Card.

During the experiment, the target stimuli were presented at 55 dB SPL (A weighted), and the level of the masker was adjusted during both procedures to converge to the SRT. For the SWN and CT condition, target and masker stimuli were presented to the right ear of the participant. For the CTBabble condition, target and masker stimuli were presented to both ears but were filtered to simulate spatial hearing. Spatial hearing was simulated using head-related transfer functions derived from measurements in an anechoic chamber (Kayser et al., 2009). This allowed us to simulate the CT to come from 90° to the left of the participant (−90°) and the target speaker from 90° to the right of the participant. The babble noise was built by first combining speech signals of 36 different speakers into 9 different babble sources. Each babble source consisted of four speech signals from two male and two female speakers. The spectra of the babble sources were separately matched to those of the speech materials. Using the head-related transfer functions, the babble sources were simulated to be present at nine equidistant positions around the participant (at −180°, −140°, . . . to 140°) and separated by 40° each.

Procedure

To obtain the SRT, both recall and rate procedure were performed by adaptively adjusting the level of the masker. This is more time efficient compared with a constant procedure as only one list is needed to obtain the SRT, instead of several at different SNRs. Furthermore, the results of a pilot study on three participants showed that a constant procedure resulted in similar SRTs, but the inter- and intrasubject variability appeared to be higher compared with the adaptive procedure. Almost all

participants completed every condition of the recall (standard adaptive) and rate (self-assessed Békesy) procedure. Only three participants did not finish the complete protocol due to a mistake of one of the experimenters. For these three subjects, one or three of 24 conditions was considered as missing data. As shown in Table 1 and Figure 1, each participant completed two procedures, the recall and rate procedure, and three blocks representing the three masker conditions: SWN, CT, and CTBabble. Both the recall and rate procedure were administered before the next masker was presented. The order of maskers was randomized across participants.

Recall procedure. Participants always started with the recall procedure where sentences were presented trial-by-trial. First, training lists were administered in order to familiarize the participant with the procedure and avoid learning effects. The adaptive procedure of Brand and Kollmeier (2002) was used for the Matrix sentences (Luts et al., 2014). The SRT of the Matrix sentences was defined as the last SNR presented in a list of 20 sentences. For the LIST sentences, the level of the masker was adjusted with a step size of 2 dB using a one-up-one-down procedure to target the SRT. The SRT was calculated by averaging the last 6 SNRs of a list of 10 sentences (van Wieringen & Wouters, 2008). The order of presenting the LIST or Matrix sentences was randomized across participants. In addition to recalling the sentences, the participants were also informed about the goal of the procedure, that is, converging to 50% speech intelligibility. By sharing this information with the participants, we aimed to train the participants in rating their speech intelligibility so they could develop a rating strategy. The underlying basis of their strategy, that is, *accurately perceiving 50% of the words* or *understanding 50% of the message,* was not explicitly questioned or documented.

Rate procedure. After presenting both Matrix and LIST sentences using the recall procedure, we administered the self-assesed Békesy procedure for the same masker. The order of presenting the different speech materials (LIST, Matrix, Story1, and Story2) was again randomized across participants. In this procedure, instead of recalling, the participants were instructed to listen to the sentences or stories and rate if their speech intelligibility was higher or lower than 50%. For the target stimuli Matrix and LIST, the sentences were concatenated into trials of approximately 2 min each (i.e., 2 lists of each 20 sentences at 2–3 s per sentence) and presented at a fixed SNR. On listening, participants could take their time and decide at any moment which rating (*<50%* or *>50%*) they wanted to give to the trial. In other words, participants did not have to rate their speech intelligibility after every sentence but instead could listen to several sentences and consciously rate at their own pace their speech intelligibility.

After rating their speech intelligibility by pushing one of the two buttons, the sentences from the current trial were immediately stopped, the SNR was adapted based on the response and the next trial, with a different set of concatenated sentences, was initiated. The level of the masker was increased if the participant pushed the button *>50%* or decreased if the *<50%* button was pushed. The initial step size was 5 dB. From the fourth trial, we changed the step size for three trials to 2 dB and to 1 dB for the remaining trials. For the stories, the method was the same except that different 1-min sections of the stories were presented per trial, not concatenated sentences. In Figure 2, the difference between the recall and rate procedure, with regard to the timeline, is shown.

The procedure was stopped when the following criterion was reached. The participant had to perform reversals in a sequential order which was indicated by a minimum sequence of the following button presses: *>50%, <50%, >50%* or vice versa *>50%, <50%, >50%*. When participants indicated that they understood 50% of the sentences or story, the procedure was ended only when a *reversal* was already performed. This method was chosen to ensure that listeners varied the SNR at least two times instead of simply ending the procedure at the first SNR without comparison. This way the experimenter guided the procedure to behave similarly to the adaptive track of the recall procedure, that is, a staircase which eventually fluctuates around the SRT. The last SNR presented was taken as the SRT of that list.

The rate procedure was conducted at least two times. If the outcome differed more than 2 dB with the previous, a third run was performed. This was done consistently, except for 6 of the 136 conditions where the experimenter did not repeat the procedure a third time while the test–retest difference was 3 or 4 dB. For our analysis, the rate SRT was calculated by averaging the last six SNRs of a list, similar to the recall SRT of the LIST sentences. By including more data points, we expected our results to be more reliable compared with only including the values of the last SNR. If three runs were administered because the test–retest criterion was exceeded, the SRT values of the last two lists were used in order to exclude learning effects and to guarantee that the participant correctly understood the instruction of the experimenter. Furthermore, the average of these values was taken to compare conditions.

Statistical Analysis

Version 3.4.4 of R was used to conduct the statistical analyses. To address the several research aims, we used

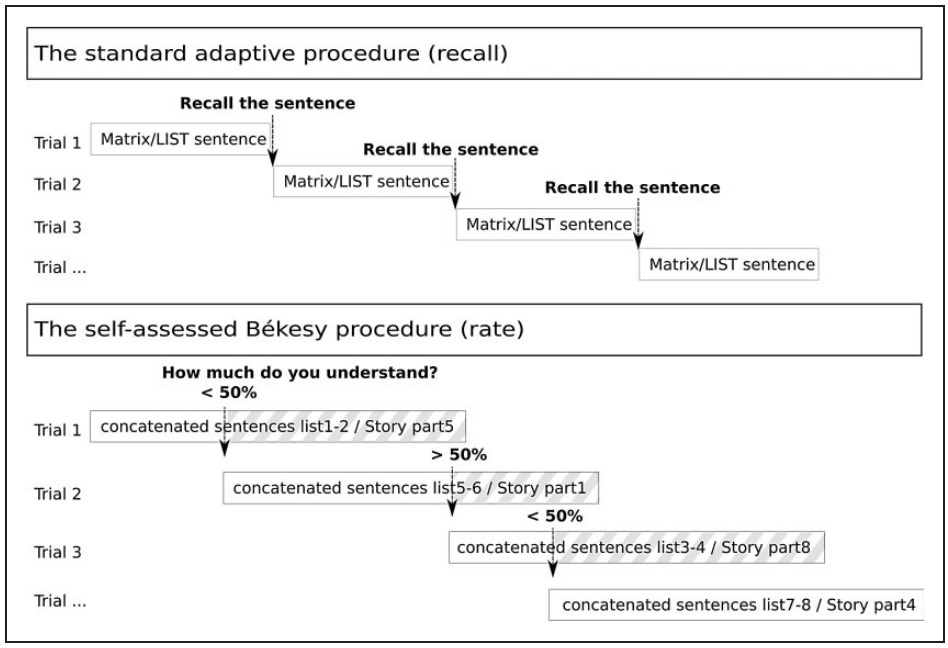

Figure 2. Timeline of the standard adaptive (recall) procedure and self-assessed Békesy (rate) procedure. During the recall procedure, participants were instructed to recall a Matrix or LIST sentence immediately after the sentence was presented. For the rate procedure, the Matrix and LIST sentences were concatenated into trials of ± 2 min each and presented at a fixed SNR. On listening, participants did not have to rate their speech intelligibility after every sentence but instead could listen to several sentences and consciously rate at their own pace by pressing one of the two buttons (*<50%* or *>50%*) at any moment. After rating, the sentences from the current trial were immediately stopped, the SNR was adapted and the next trial was initiated. For the stories, the method was the same except that different 1-min sections of the stories were presented per trial.
Note. LIST = Leuven intelligibility sentence test.

nonparametric tests since we recruited a small sample size of 14 participants for the first evaluation of our new procedure. Since the distribution of small data sets is difficult to examine, it is often recommended to use nonparametric tests as these do not have assumptions about the distribution of the data (Altman, Gore, Gardner, & Pocock, 1992). In addition to this, Shapiro–Wilk test ($p < .05$) revealed a significant deviation from the normal distribution in at least one of the conditions. Wilcoxon signed-rank tests were chosen because the same participants completed the different conditions of interest. To investigate differences in intersubject variability between conditions, we used the Levene test centered around the median. To assess the reliability of the self-assessed Békesy procedure (rate procedure), we evaluated the intrasubject variability determined by the standard deviation of repeated measurements. This measure is frequently used when evaluating speech-in-noise tests (Jansen et al., 2012; Luts et al., 2014; Nilsson, Soli, & Sullivan, 1994; van Wieringen & Wouters, 2008; Warzybok et al., 2015). In this study, we calculated these standard deviations by taking the root mean square of the differences between test and retest, divided by $\sqrt{2}$ (Plomp & Mimpen, 1979; Wagener & Brand, 2005). Lastly, the method of Holm (1979) was applied to control for multiple comparisons if needed.

Results

Recalling Versus Rating Standardized Sentences

A first step in the validation of our procedure is investigating how the outcomes of the rate procedure are related to those of the recall procedure. In Figure 3, the rate and recall SRTs of the standardized Matrix and LIST sentences are compared. First, it can be seen that the rate SRTs are closer to the recall SRTs for the Matrix compared with the LIST. No significant difference was found between the rate and recall SRT of the Matrix sentences for all masker conditions, using Wilcoxon signed-rank tests. Median differences of −0.5, −0.6, and −3.3 dB between the rate and recall Matrix SRTs were observed for SWN, CTBabble, and CT, respectively. In contrast to the Matrix, most SRTs of the LIST sentences are situated above the diagonal. In line with this, Wilcoxon signed-rank tests and the median of the intraindividual differences showed significantly lower rate than recall SRTs for the LIST sentences, in all the masker conditions (SWN: median diff = −3.7 dB,

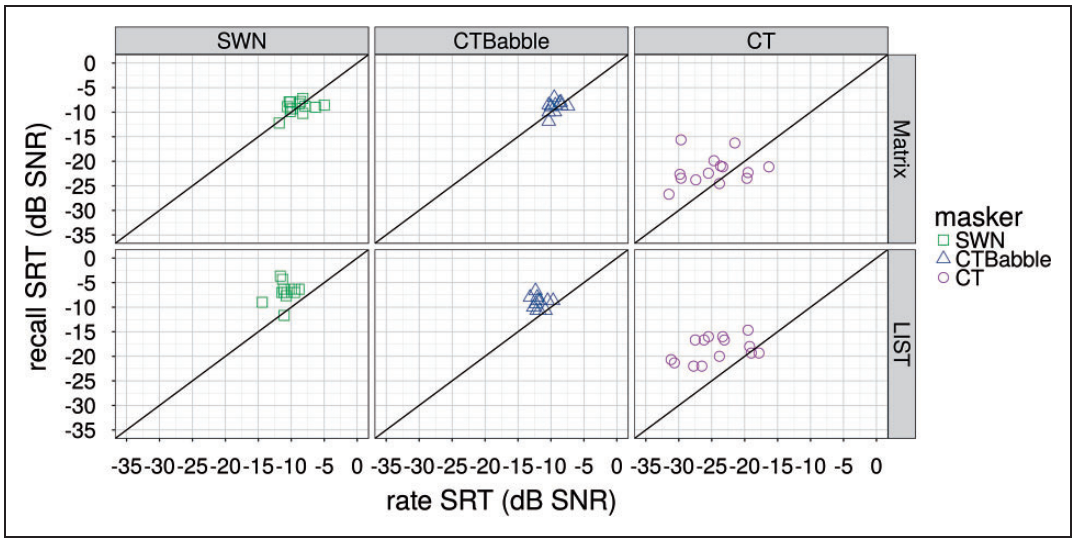

Figure 3. Scatterplot showing the recall SRTs of the Matrix and LIST sentences against their rate SRTs. The symbols and colors represent the three masker conditions: SWN, CTBabble, and CT.

Note. SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise; SRT = speech reception threshold; SNR = signal-to-noise ratio.

$p < .001$; CT: median diff = −6.1 dB, $p = .001$; CTBabble: median diff = −2.7 dB, $p < .001$).

Inter- and Intrasubject Variability

Intersubject variability. Figure 4 shows the SRTs for both the recall and rate procedure for all speech materials and maskers. With regard to the effect of speech material on the intersubject variability, the boxplots show similar interquartile ranges for the sentences and connected discourse materials. Using Levene's test, we found no significant differences between the intersubject variabilities of the different speech materials for any masker type with either of the procedures.

However, an increase is seen in interquartile range for the sentences when participants are instructed to rate their speech intelligibility compared with recalling the sentences. Using Levene's test, only when CT was used as masker, a significant increase was detected between the variances of recall and rate LIST SRTs ($F(1, 26) = 4.285$, $p = .049$). For the Matrix sentences or other masker conditions, no significant differences in variances between procedures were found.

With regard to the effect of masker, Figures 3 and 4 clearly show an increase in intersubject variability with CT compared with the other two maskers (SWN and CTBabble). Using Holm-adjusted Levene's test, we found a significant increase in intersubject variability of the rate SRTs when CT is used as masker compared with SWN or CTBabble for all speech materials (Matrix, LIST, and DWZ: $p < .05$). For the recall procedure, only for the LIST sentences a significant increase in intersubject variability was detected for CT compared with CTBabble ($F(1, 26) = 10.384$, $p = .01$).

Intrasubject variability. To check if a procedure is reliable, it is important to assess the intrasubject variability or test–retest reliability. Table 2 summarizes the standard deviations of the repeated measurements for the different conditions. We can infer from this table that a relatively small test–retest variability was obtained for all speech materials in SWN and CTBabble (range: 0.7–1.4 dB), except for Story1 (2 dB). Furthermore, a substantially higher intrasubject variability was obtained for all speech materials when CT was used as masker (range: 1.6–2.2 dB).

Effect of Masker Type on the Recall and Rate SRTs

With regard to the effect of masker, we found significantly lower (better) rate and recall SRTs when using CT compared with SWN or CTBabble (Figure 4 and Holm-adjusted Wilcoxon signed-rank tests: $p < .01$). The large positive values reported in Table 3 confirm this effect for all speech materials and both procedures. In contrast with this, smaller differences between SWN and CTBabble were found. Wilcoxon signed-rank tests only detected a significantly lower recall LIST SRT when comparing CTBabble to SWN (median diff = −2.3 dB; $p = .008$).

In addition to this, we investigated if the recall and rate procedure led to similar effects of masker type, described earlier. In other words, lowest SRTs when CT is used as masker and similar SRTs for SWN and

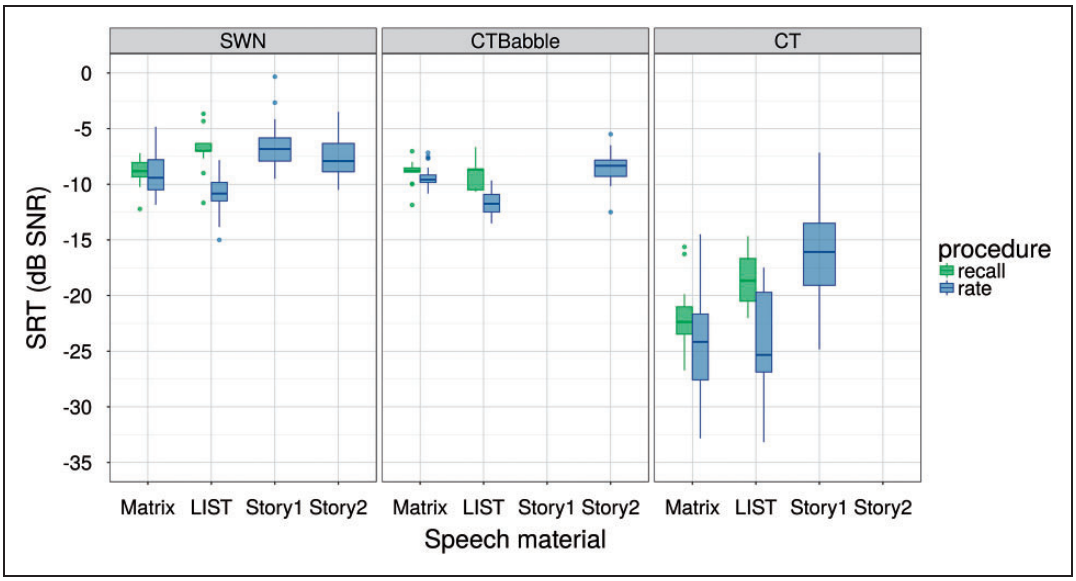

Figure 4. Boxplots of the recall and the rate SRTs of the four speech materials: Matrix, LIST sentences, Story1, and Story2 in the three maskers: SWN, CTBabble, and CT.
Note. SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise; SRT = speech reception threshold; SNR = signal-to-noise ratio.

CTBabble. Using, Holm-adjusted Wilcoxon signed-rank tests, we found no significant differences between the recall and rate procedure, except for the LIST sentences when comparing CT versus CTBabble (Table 3; $p = .018$). In line with this, confidence intervals of the difference between recall and rate procedure were narrow and close to zero when comparing SWN with CTBabble. When comparing SWN or CTBabble to CT, however, wider intervals were obtained. This may be due to the slightly larger differences between CT and the other maskers for the rate compared with recall procedure.

Lastly, we can infer from Table 3 that a similar trend of masker type was found on the intelligibility of connected discourse compared with the sentences. In other words, a higher Story1 SRT was found for SWN compared with CT (median diff = 9.2 dB), and a similar Story2 SRT was found for SWN versus CTBabble (median diff = 0.7 dB). Although the trend is similar, we have to note that the difference between SWN and CT is substantially smaller (median diff = 9.2 dB) compared with the Matrix (median diff = 16.5 dB) and LIST sentences (median diff = 14.4 dB).

Sentences Versus Connected Discourse

Since we developed a new method to measure intelligibility of connected discourse, it is interesting to study how intelligibility of connected discourse relates to that of sentences. As shown in Table 4 and Figure 4, our data reveal that rate SRTs for Story1 and Story2 were significantly higher compared with those of the sentences for all masker conditions ($p < .05$). Furthermore, the negative differences between the rate SRTs of the sentences and Story1 and the positive but very small difference between Story1 and Story2 indicate that Story1 was the most difficult to understand. When taking the size of the differences into account, differences which may influence the interpretation of the outcomes of experiments were primarily obtained in SWN and CTBabble when comparing the LIST sentences with Story1 (SWN: −3.7 dB) or Story2 (SWN: −3 dB and CTBabble: −3.4 dB; Table 4). In addition, when CT was used as masker, a large difference of approximately −9 dB was found when comparing both LIST or Matrix sentences with Story1.

Table 2. Test–Retest Variability of the Rate Procedure Determined by the Standard Deviation of the Repeated Measures (Plomp & Mimpen, 1979; Wagener & Brand, 2005).

	SWN	CTBabble	CT
Matrix	0.7 dB	0.7 dB	1.9 dB
LIST	1.1 dB	0.8 dB	1.6 dB
Story1	2 dB		2.2 dB
Story2	1.4 dB	0.8 dB	

Note. SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise; LIST = Leuven intelligibility sentence test.

Table 3. Median Values of the Intraindividual Differences Between Masker Conditions for All Speech Materials in Both Procedures.

		Recall	Rate	*p*	CI of difference
SWN vs. CT	Matrix	12.8 dB	16.5 dB	.051	[0.3, 6.1]
	LIST	11.7 dB	14.4 dB	.115	[−0.3, 5]
	Story1		9.2 dB		
SWN vs. CTBabble	Matrix	0 dB	−0.2 dB	.735	[−0.9, 1.6]
	LIST	2.3 dB	1.3 dB	.115	[−2.7, 0]
	Story2		0.7 dB		
CTBabble vs. CT	Matrix	13.5 dB	15.5 dB	.065	[0.2, 5.8]
	LIST	9.3 dB	13.2 dB	.018	[1.6, 5.7]

Note. Holm-adjusted Wilcoxon signed-rank tests were used to investigate if the effect of masker type on sentences differed depending on the procedure being used (recall vs. rate). The *p* values and confidence intervals of the differences between the rate and recall procedure are reported (significance level = .05).
SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise; LIST = Leuven intelligibility sentence test.

Table 4. Median Values of the Intraindividual Differences Between the SRTs of the Speech Material Reported in the Row Versus Column.

		Story1	Story2
SWN	Matrix	−2.1 dB ($p = .001$)	−1.7 dB ($p = .017$)
	LIST	−3.7 dB ($p = .001$)	−3 dB ($p = .001$)
	Story1		0.8 dB ($p = .017$)
CT	Matrix	−10 dB ($p = .001$)	
	LIST	−8.2 dB ($p = .001$)	
CTBabble	Matrix		−0.8 dB ($p = .05$)
	LIST		−3.4 dB ($p = .004$)

Note. For example, a negative value (e.g., −2.1 dB) means that we have obtained a lower (better) SRT for the Matrix sentences compared with Story1. Holm-adjusted Wilcoxon signed-rank tests were used to investigate if the difference between speech materials was significant. The *p* values are reported between brackets (significance level = .05).
SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise; LIST = Leuven intelligibility sentence test.

Time Efficiency

To implement this method in research and the clinic, it is important that our new procedure does not require a long test duration. By using an adaptive instead of a constant procedure, the SRT of an individual is obtained during only one run of the procedure. The test durations summarized in Figure 5 show that our rate procedure takes substantially less time compared with the recall procedure of the Matrix sentences whereas a similar test duration was found to the one of the LIST. This can be explained by the different number of Matrix (20) and LIST (10) sentences that have to be recalled during one run. Across maskers and speech materials, we can infer from Figure 5 that the test durations ranged from ± 1 min to ± 6 min due to the long test durations of the recall procedure of the Matrix sentences. When we exclude the latter procedure, the test duration ranged between 1 and 2 min, with interquartile range between 80 and 130 s. In addition to this, our data show that most participants needed between 10 and 16 trials during the rate procedure to converge to the SRT.

Discussion

In the present study, we described and evaluated the self-assessed Békesy procedure, a new method that is able to measure intelligibility of sentences and connected discourse. In young, normal-hearing participants, our data show that our method is a valid, reliable, and time efficient procedure. For both recall and rate procedure, better SRTs were found for sentences and connected discourse when CT was used as masker compared with SWN and CTBabble. This indicates the ability of the self-assessed Békesy procedure to detect similar effects of masker type compared with a recall procedure as well as to investigate the effect of masker type on the intelligibility of connected discourse. In addition, significant differences between the rate SRTs of LIST sentences versus connected discourse materials were found for all masker conditions.

Recalling Versus Rating Standardized Sentences

Our data show small median differences (< 1 dB) between the rate and recall SRTs of the Matrix sentences when SWN and CTBabble were used as masker. This indicates the feasibility of the self-assessed Békesy procedure to be implemented in research. When using CT as masker, a slightly higher, but nonsignificant difference of 3.3 dB was found. For the LIST sentences, however, significantly better rate than recall SRTs were found for all masker conditions (range of median differences: 2.7–6.1 dB). We hypothesize that the difference between scoring method and rating strategy can explain these results (Falconer & Davis, 1947).

First, multiple participants reported that the rating of the Matrix sentences was easier compared with the LIST because the fixed Matrix structure of five words per sentence allowed them to count (Luts et al., 2014). This consequently resulted in a more objective estimation which may explain the similar Matrix recall and rate SRTs. Because of the fixed structure and low context information, we believe that Matrix rate SRTs capture more information about accurately perceiving words rather than understanding the message. In contrast to this, LIST sentences contain more context and resemble more

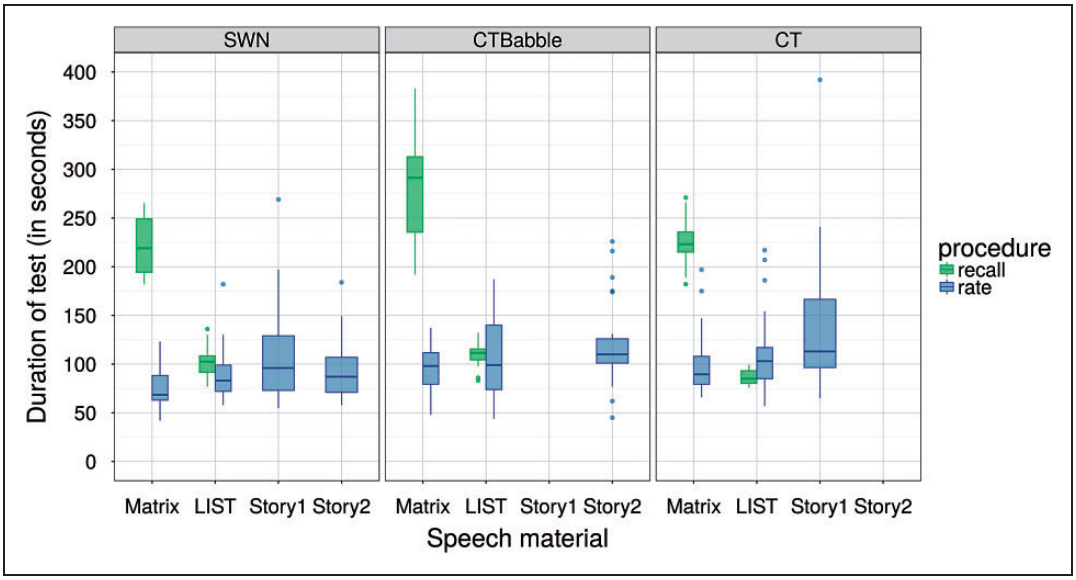

Figure 5. Boxplots showing the duration of the recall and rate procedure (in seconds), conducted for all speech materials and masker types. The two procedures, recall and rate, are coded in color.
Note. SWN = speech weighted noise; CT = competing talker; CTBabble = competing talker in combination with babble noise.

connected discourse and daily life sentences. Although, the scoring method of the LIST sentences is based on keywords, all keywords should be recalled correctly in order to score the sentence correct (van Wieringen & Wouters, 2008). In other words, when only four of five keywords are recalled correctly, the LIST sentence is scored incorrect during the recall procedure while these sentences are probably rated as highly intelligible. As participants are not informed about the keyword-based scoring method, we expect that this difference between scoring method and rating strategy led to the difference between procedures. In view of speech intelligibility, we could state that the recall LIST SRTs reflect accurately perceiving the words while during the rating procedure the understanding of the message was captured.

To have a better comparison between these procedures in the future, the scoring method of the LIST sentences could be adjusted to the number of correct keywords to adapt the SNR. This way, the degree of understanding the message may be better approximated. In addition to this, making the instructions of the experimenter more explicit or documenting the rating strategy could allow to assess how persons rate their speech intelligibility, that is, based on perceiving words or understanding the message.

Inter- and Intrasubject Variability

By investigating the intersubject variability, we could study the discriminative power of our procedure but also the variation related to procedural and stimulus aspects. Based on the literature, we assumed that individual differences are better detected when presenting connected discourse instead of sentences (MacPherson & Akeroyd, 2014). However, in this study, we did not find a significant effect of speech material on the intersubject variability. While MacPherson and Akeroyd (2014) recruited older listeners with different degrees of hearing loss, our participant group was very homogeneous as it consisted only of young, normal-hearing participants. Despite this, we found higher intersubject variabilities for our fluctuating noise (CT) which is consistent with validation studies of speech-in-noise tests (Francart et al., 2011; Jansen et al., 2012; Wagener & Brand, 2005). Although this could reflect a better discriminative power, it can also be related to the variation in stimulus aspects. As connected discourse passages are often heterogeneous in terms of acoustics, content as well as the length of the silent gaps, a randomization of the segments of the target and CT across participants can result in an increased intersubject variability of the SRT (Kei et al., 1999; Wagener & Brand, 2005).

To quantify the reliability of our rate procedure, we also evaluated the test–retest reliability or intrasubject variability. Although the test–retest reliability for the rate Matrix SRTs ($\pm$ 0.7 dB) is higher compared with values obtained during the Matrix recall sentence test (0.4–0.5 dB; Jansen et al., 2012; Luts et al., 2014; Warzybok et al., 2015), our test–retest reliability when SWN and CTBabble are used as masker ($<$ 1.5 dB) is similar to those found with methods measuring

intelligibility of connected discourse (Best, Keidser, Freeston, & Buchholz, 2016; Falconer & Davis, 1947) and other speech-in-noise tests using HINT or LIST sentences (Jansen et al., 2012; Nilsson et al., 1994; van Wieringen & Wouters, 2008). In addition to this, a substantially larger test–retest reliability (± 2 dB) was found when Story1 or CT were presented as target stimulus or masker, respectively. Similar to other studies, stories or CTs are often heterogeneous in terms of acoustics and length of the silent gaps (Kei et al., 1999; Wagener & Brand, 2005). To control for this in further research, we could limit the silences of both target talker and masker to 200 ms, which will preserve the naturalness of the speech more compared with gaps of approximately 100 ms.

Effect of Masker Type and Speech Material

In this study, we used three maskers to verify if the self-assessed Békesy procedure detects similar effects of masker type compared with the recall procedure. This is valuable in the sense that studies have shown that older adults experience more difficulties with informational maskers compared with energetic (Goossens et al., 2017). Similar to the literature, our data showed significantly better SRTs for young, normal-hearing participants for both procedures when CT was used as masker compared with SWN or CTBabble (Francart et al., 2011; Goossens et al., 2017). In other words, both the recall and rate procedure reliably detected the benefit of temporal gaps for speech intelligibility. For the most realistic masker, CTBabble, similar SRTs were found to SWN with either of the procedures. Although slightly larger differences between the maskers were observed for the rate procedure, similar conclusions could be made for the effect of masker type on the recall and rate SRTs. To the best of our knowledge, this is also the first study which has compared different masker types on the intelligibility of connected discourse. Although a similar trend was found compared with the sentences, the smaller difference in SRT between SWN and CT for Story1 compared with LIST and Matrix (Table 3) suggests that our participants benefited less from a CT when the target stimulus was more complex. More studies using connected discourse tests have to be conducted to confirm this.

Lastly, our data show that it is important to take differences between continuous speech (i.e., sentences) and connected discourse into account. Significantly, higher SRTs were obtained for connected discourse compared with the Matrix and LIST sentences, especially when CT was used as masker (Matrix/LIST vs. Story1: −9 dB). This is in line with previous results of MacPherson and Akeroyd (2014) who also found worse SRTs for connected discourse compared with continuous speech (i.e., sentences) in SWN. As a result of this, it seems that connected discourse materials should be carefully selected because an effect of interest may in reality reflect differences in content or acoustics between- or within-speech materials. For this reason, we advise for future behavioral or electrophysiological tests to use the outcomes of the self-assessed Békesy procedure, as a measure for the SRT of connected discourse instead of presenting these materials at the same SNRs as standardized sentences. In other words, norms can be established for connected discourse, using standardized sentences as a reference.

Future Work

Although our self-assessed Békesy procedure has a good time efficiency, more steps have to be taken before implementing this new procedure in research and the clinic. First of all, we noticed that participants found it difficult to use the buttons *>50%* and *<50%* to indicate their level of speech intelligibility. Therefore, we suggest to implement a continuous scale from 0% to 100%. Similar to the buttons, this scale will motivate the participants to consciously rate their speech intelligibility but will also allow them to estimate it in a more natural way. In addition, we can use this extra information to fit psychometric functions on our data. A second important step is to validate this new procedure in a more heterogenous group such as older individuals with different degrees of hearing loss. This not only allows us to investigate if the self-assessed Békesy procedure is sensitive to individual differences but also to examine if our rate procedure is still time efficient and reliable. Older persons are known to be slower, often underestimate their performance and need more effort to perform in the same way (Gosselin & Gagné, 2011). These factors may affect the rating scores and consequently result in differences with the outcomes on speech audiometry tests. Although this may seem problematic, capturing differences in effort can also mean that the self-assessed Békesy procedure resembles daily life performance more closely.

Although the primary goal of our study was to find a link with the outcomes of current speech audiometry tests, it should be investigated if our self-assessed Békesy procedure can also be used to get information about real-world performance. As a first step, a higher speech intelligibility level which more resembles daily life, such as 80% could be used as target point for this new procedure. Smits and Festen (2011), for example, found that the intersubject variability is lower when converging to higher speech intelligibility levels on the psychometric curve. Moreover, they showed that the maximal reliability should be achieved at approximately 80% if the slope of the psychometric curve is shallow. Although it is hard to imagine what 80% sounds like, we hypothesize that

implementing a scale (0%–100%) instead of the two buttons (>50% or <50%), as mentioned earlier, could solve this problem. This way, participants do not have to focus on 80% but can compare between different trials. In addition to a new target point, other measures important for understanding speech in daily life could be related with the outcomes of our self-assessed Békesy procedure. For example, outcomes on cognitive tests such as the reading span test (Best, Keidser, Freeston, et al., 2016) or questionnaires about speech-in-noise performance and listening effort (Gatehouse & Noble, 2004; Gosselin & Gagné 2011) could be linked to the individual SRTs when testing a more heterogenous group of older hearing impaired persons. Finally, it would also be interesting to test if the self-assessed Békesy procedure can be used to evaluate the benefit of hearing aids and their different processing schemes while presenting connected discourse.

Conclusion

In this study, we proposed a new method, the self-assessed Békesy procedure, to determine the SRT of connected discourse. When the scoring method of the recall procedure is comparable to the strategy assumed to underlie the person's ratings, we obtained similar results for both procedures. This suggests that the self-assessed Békesy procedure is a valid procedure that can be used in an experimental setting. In general, a good test–retest reliability (< 1.5 dB) comparable to the standard speech-in-noise tests was obtained. Only when CT was used as masker, higher inter- and intrasubject variabilities were observed which could be due to variations in stimulus aspects. Furthermore, similar effects of masker types were found for recall and rate procedure which supports the potential of our procedure to investigate the effect of masker type in young versus older listeners. In addition to this, our results suggest that intelligibility of connected discourse may be differently affected by the masker type compared with sentences. Finally, the differences between the LIST sentences and connected discourse materials indicate the importance of controlling for differences in intelligibility between sentences and connected discourse.

Acknowledgments

The authors are very grateful to all participants for their participation in our study as well as to the master's students Elien Van den Borre and Jelien Geivers for their help in data acquisition. The authors also would like to thank Michael Akeroyd (Associate Editor) and the two anonymous reviewers for the comments on the manuscript.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has received funding from the Europe project and Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Number 637424). Furthermore, financial support was provided by the KU Leuven Special Research Fund to Tom Francart (Grant Number OT/14/119). Lastly, research of Eline Verschueren is funded by a PhD grant of the Research Foundation Flanders (FWO; Grant Number 1S86118N).

ORCID iD

Lien Decruy http://orcid.org/0000-0002-2983-9972

References

Altman, D. G., Gore, S. M., Gardner, M. J., & Pocock, S. M. (1992). Statistical guidelines for contribution to medical journals. *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, *29*, 1–8. doi:0.1177/000456329202900101.

Anderson-Gosselin, P., & Gagne, J. (2010). Use of a dual-task paradigm to measure listening effort. *Canadian Journal of Speech-Language Pathology and Audiology/Revue canadienne d 'orthophonie et d'audiologie*, *34*(1), 43–51.

Best, V., Keidser, G., Buchholz, J. M., & Freeston, K. (2016). Development and preliminary evaluation of a new test of ongoing speech comprehension. *International Journal of Audiology*, *55*(1), 45–52. doi:10.3109/14992027.2015.1055835

Best, V., Keidser, G., Freeston, K., & Buchholz, J. M. (2016). A dynamic speech comprehension test for assessing real-world listening ability. *Journal of the American Academy of Audiology*, *27*(7), 515–526. doi:10.3766/jaaa.15089

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer* [Computer program]. Retrieved from http://www.praat.org

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

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735. doi:10.1523/JNEUROSCI.5297-12.2013

Falconer, G., & Davis, H. (1947). The intelligibility of connected discourse as a test for the threshold for speech. *The Laryngoscope*, *57*(9), 581–595.

Francart, T., van Wieringen, A., & Wouters, J. (2008). APEX 3: A multi-purpose test platform for auditory psychophysical experiments. *Journal of Neuroscience Methods*, *172*(2), 283–293. doi:10.1016/j.jneumeth.2008.04.020

Francart, T., van Wieringen, A., & Wouters, J. (2011). Comparison of fluctuating maskers for speech recognition tests. *International Journal of Audiology*, *50*(1), 2–13. doi:10.3109/14992027.2010.505582

Gatehouse, S., & Noble, W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology*, *43*(2), 85–99. doi:10.1080/14992020400050014

Goossens, T., Vercammen, C., Wouters, J., & van Wieringen, A. (2017). Masked speech perception across the adult lifespan: Impact of age and hearing impairment. *Hearing Research*, *344*, 109–124. doi:10.1016/j.heares.2016.11.004

Gosselin, A. P., & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *International Journal of Audiology*, *54*, 944–958. doi:10.1044/1092-4388(2010/10-0069)

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70. doi:10.2307/4615733

Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R., . . . Van Wieringen, A. (2012). Comparison of three types of French speech-in-noise tests: A multicenter study. *International Journal of Audiology*, *51*(3), 164–173. doi:10.3109/14992027.2011.633568

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., & Kolmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, *2009*(1), 298605. doi:10.1155/2009/298605

Kei, J., Smyth, V., Murdoch, B., & McPherson, B. (1999). Measuring the understanding of connected discourse: An overview of methodology and clinical applications in rehabilitative audiology. *Asia Pacific Journal of Speech, Language and Hearing*, *4*(1), 13–37. doi:10.1179/136132899805577169

Kidd, G., & Colburn, H. S. (2017). Informational masking in speech recognition. In J. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party* (pp. 75–109). Cham, Switzerland: Springer International Publishing.

Levitt, H. (1971). Transformed up down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. doi:10.1121/1.1912375

Luts, H., Jansen, S., Dreschler, W., & Wouters, J. (2014). *Development and normative data for the Flemish/Dutch Matrix test* (Technical report). Retrieved from https://lirias.kuleuven.be/bitstream/123456789/474335/1/Documentation+Flemish-Dutch+Matrix_December2014.pdf

MacPherson, A., & Akeroyd, M. A. (2014). A method for measuring the intelligibility of uninterrupted, continuous speech. *The Journal of the Acoustical Society of America*, *135*(3), 1027–1030. doi:10.1121/1.4863657

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

Petersen, E. B., Wöstmann, M., Obleser, J., & Lunner, T. (2017). Neural tracking of attended versus ignored speech is differentially affected by hearing loss. *Journal of Neurophysiology*, *117*(1), 18–27. doi:10.1152/jn.00527.2016

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

Smits, C., & Festen, J. M. (2011). The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: Steady-state noise. *The Journal of the Acoustical Society of America*, *130*(5), 2987–2998. doi:10.1121/1.3644909

Speaks, C., Trine, T. D., Crain, T. R., & Niccum, N. (1994). A revised speech intelligibility rating (RSIR) test: Listeners with normal hearing. *Otolaryngology-Head and Neck Surgery*, *110*(1), 75–83. doi:10.1177/019459989411000109

van Wieringen, A., & Wouters, J. (2008). LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. *International Journal of Audiology*, *47*(6), 348–355. doi:10.1080/14992020801895144

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

Walker, G., & Byrne, D. (1985). Reliability of speech intelligibility estimation for measuring speech reception thresholds in quiet and in noise. *The Australian Journal of Audiology*, *7*, 23–31.

Warzybok, A., Zokoll, M., Wardenga, N., Ozimek, E., Boboshko, M., & Kolmeier, B. (2015). Development of the Russian matrix sentence test. *International Journal of Audiology*, *54*, 35–43. doi:10.3109/14992027.2015.1020969

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. doi:10.1097/AUD.0b013e3181d4f251