Guided ecological momentary assessment in real and virtual sound environments

Mansour, Naim; Westermann, Adam; Marschall, Marton; May, Tobias; Dau, Torsten; Buchholz, Jörg

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/10.0006568

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Guided ecological momentary assessment in real and virtual sound environments

Naim Mansour, Adam Westermann, Marton Marschall, et al.

Citation: The Journal of the Acoustical Society of America **150**, 2695 (2021); doi: 10.1121/10.0006568
View online: https://doi.org/10.1121/10.0006568
View Table of Contents: https://asa.scitation.org/toc/jas/150/4
Published by the Acoustical Society of America

JASA ARTICLE

Guided ecological momentary assessment in real and virtual sound environments

Naim Mansour,[1,a)] Adam Westermann,[2] Marton Marschall,[1,b)] Tobias May,[1] Torsten Dau,[1,c)] and Jörg Buchholz[3,d)]
[1]*Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark*
[2]*Widex A/S, Lynge, Denmark*
[3]*Department of Linguistics and Audiology, Macquarie University, Sydney, Australia*

ABSTRACT:
Ecological momentary assessment (EMA) outcome measures can relate people's subjective auditory experience to their objective acoustical reality. While highly realistic, EMA data often contain considerable variability, such that it can be difficult to interpret the results with respect to differences in people's hearing ability. To address this challenge, a method for "guided" EMA is proposed and evaluated. Accompanied and instructed by a guide, normal-hearing participants carried out specific passive and active listening tasks inside a real-world public lunch scenario and answered EMA questionnaires related to aspects of spatial hearing, listening ability, quality, and effort. *In situ* speech and background noise levels were tracked, allowing the guided EMA task to be repeated inside two acoustically matched, loudspeaker-based laboratory environments: a 64-channel virtual sound environment (VSE) and a three-channel audiology clinic setup. Results showed that guided EMA provided consistent passive listening assessments across participants and conditions. During active listening, the clinic setup was found to be less challenging than the real-world and the VSE conditions. The proposed guided EMA approach may provide more focused real-world assessments and can be applied in realistic laboratory settings to aid the development of ecologically valid hearing testing.
 https://doi.org/10.1121/10.0006568
(Received 12 March 2021; revised 27 August 2021; accepted 17 September 2021; published online 12 October 2021)
[Editor: Jonas Braasch] Pages: 2695–2704

I. INTRODUCTION

Relating the subjective auditory experience of an individual person in complex, real-world environments to objective measures of hearing ability has been of interest within hearing research for many years. Most studies have focused on evaluating psychoacoustic measures like loudness perception, spatial awareness, speech intelligibility, and localization ability using well-controlled, yet artificial stimuli. Efforts have been made to increase the ecological validity (Reis and Judd, 2000) of these stimuli by reproducing real-world environments in the laboratory (e.g., Ahrens *et al.*, 2017; Best *et al.*, 2015; Mansour *et al.*, 2019b; Westermann and Buchholz, 2015). However, it is unclear how the task paradigms in such studies and their corresponding outcome measures, though reliable and reproducible, can be related to subjective hearing ability in the real world (Lutman, 1991; Timmer *et al.*, 2015). Retrospective questionnaires—like the Speech, Spatial and Qualities of Hearing Scale (SSQ) (Gatehouse and Noble, 2004) or the Glasgow Hearing Aid Benefit Profile (GHABP) (Gatehouse, 1999)—were developed specifically to quantify subjective hearing ability. While their responses can correlate well with objective measures, e.g., better-ear average hearing thresholds, (Gatehouse and Noble, 2004), providing so-called construct validity, responses are often affected by recall bias (Moskowitz and Young, 2006), limiting their reliability.

To overcome these issues, the methodology of ecological momentary assessment (EMA) has emerged as an approach in which questionnaires are employed to capture people's subjective environmental impressions at frequent (chosen or triggered) intervals over an extended period of time in their every-day life (Shiffman *et al.*, 2008). Effects of recall bias can largely be avoided since participants evaluate their surroundings while they are observing them and instrumental measures (e.g., background noise levels and frequency spectra) can be applied to acoustically characterize the *in situ* environment. Several recent studies have applied the methodology of EMA to hearing research, establishing its reliability and construct validity (Galvez *et al.*, 2012; Henry *et al.*, 2012; Timmer *et al.*, 2017; Wu *et al.*, 2015).

However, drawbacks to the successful use of EMA remain. As reported in Timmer *et al.* (2017), there are potential issues of compliance (the participant's willingness to complete the assessments), feasibility (the extent to which the participant can fulfill the EMA task requirements), burden (the demand placed on the participant) and data variability (the large inter-subject variability contained within

[a)]Electronic mail: naiman@dtu.dk, ORCID: 0000-0001-5673-6840.
[b)]ORCID: 0000-0003-2534-7062.
[c)]ORCID: 0000-0001-8110-4343.
[d)]ORCID: 0000-0001-6188-9761.

EMA data sets). Since these issues are inherent to the EMA methodology, all previous studies suffer from them. The variability in EMA data has two main sources: temporal, or intra-participant variability, i.e., the spread within a single participant's questionnaire responses over time due to the changing environment and circumstances, and inter-participant variability, i.e., the spread across different participants' responses due to differences in their everyday environment and their hearing ability. Intra-participant variability may be further affected by changes in the participant's mood, while inter-participant variability might be impacted by individual differences in loudness assessment, speech intelligibility, and so on. While EMA data in their traditional form are useful for characterizing experienced trends and differences across large groups of people and data sets, an approach with reduced intra- and inter-participant variability may be beneficial to explore measures of individual hearing ability.

In this exploratory study, a "guided" approach to EMA is proposed, which attempts to improve participant compliance and task feasibility, lessen the burden placed on the participant, and in particular, reduce the variability in the EMA data. During the experiment, which took place at a fixed time of the day and over a short time interval, each participant visited the same, predetermined real-world (RW) scene, accompanied by a human guide. The location and time-of-day constraints served to reduce the inter-participant variability in the EMA data, while the duration constraint aimed to reduce the intra-participant variability caused by environmental changes. The guide facilitated the participant's EMAs, attempting to improve compliance and feasibility and reduce burden, by providing a clearly structured passive listening task, a communication task, and an active listening task. Each task was followed by a brief questionnaire on commonly addressed hearing topics in EMA research. The passive listening task required the participant to simply listen to their environment for 1 min, while the conversation task consisted of 1 min of natural conversation between the participant and the guide. During the active listening task, the participant was asked to listen to two monologues told by the guide, the first held at natural speech levels, established during the communication task, and the second at challenging speech levels, subjectively determined by the guide. These tasks were designed to mimic common aspects of unguided EMAs in a more controlled fashion, while the inclusion of the active listening task at challenging speech levels was intended to avoid potential ceiling effects.

Due to its rigid structure and short duration, the guided EMA method could be applied inside realistic laboratory environments, represented by a virtual sound environment (VSE) in the form of a spherical 64-loudspeaker array and a three-loudspeaker (front-left-right) clinic environment (CL); a simple, yet common, setup in audiology clinics. Both laboratory environments employed Ambisonic renderings of spatial audio recordings made in the RW environment to represent the background noise, as well as prerecorded audiovisual (AV) clips to visually represent the guide's speech during the active listening tasks. The use of these prerecorded stimuli was intended to further reduce the inter- and intra-participant variability caused by acoustic changes in the speech and noise signals in the RW environment. The background noise levels and the guide's speech levels occurring in the RW environment were matched inside the lab using an *in situ* signal-to-noise ratio (SNR) estimation method. This matching technique aimed at reducing the intra-subject variability in the EMA data between the RW environment and the laboratory environments. Both the use of guided EMA and the level matching were undertaken to compare the participants' laboratory EMA data with their real-world EMA data and to investigate the consistency of the results across participants and conditions. It was hypothesized that the EMA results for both the passive and the active listening tasks would be consistent (i.e., have a low variability) between participants in the RW condition, and that the RW EMA results would be similar to the corresponding results in the VSE and CL conditions. In this exploratory work, only normal-hearing participants were included; yet, by providing consistent results for this listener group, the guided EMA method could potentially be used to relate listeners' EMAs to their degree of hearing loss in both real-world and laboratory environments.

II. METHODS

A. Real-world assessment

The RW assessment phase of the experiment took place inside a canteen on a university campus over lunch time. Such a public lunch scenario is known to occur commonly in people's lives and is generally rated as important and challenging to hear in (Mansour *et al.*, 2019b; Wolters *et al.*, 2016). The public lunch scenario was characterized acoustically by a reverberation time (RT_{60}) of 2.5 s and an early decay time (EDT) of 0.2 s. As shown in Fig. 1, a participant and the guide were seated across from each other at a table without other occupants, with a distance of 1 m between them [resulting in a direct-to-reverberant ratio (DRR) of 7.4 dB]. Several similar tables were placed in the immediate surrounding area, populated by people having lunch. As outlined in the left column of Fig. 2, the participant, following the instructions of the guide, completed a passive listening task, a communication task, and an active listening task. In the passive listening task, the participant listened to their surroundings for 1 min without talking and then completed a 4-part EMA questionnaire using a proprietary smartphone app. The communication task consisted of a 1-min-long unscripted conversation between the participant and the guide, in order to determine the natural, conversational speech levels in the scenario. Finally, the participant listened to two 1-min monologues held by the guide; the first at the established conversational speech levels and the second at challenging levels that were subjectively determined by the guide. After each monologue, the participant answered another 4-part EMA questionnaire on their listening

https://doi.org/10.1121/10.0006568

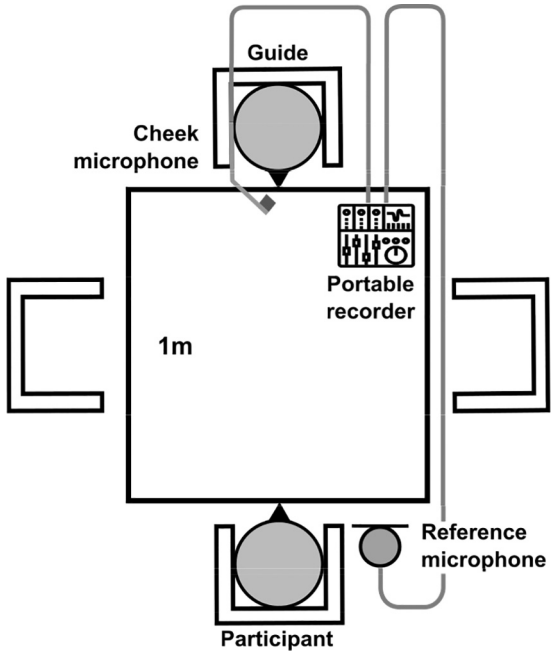

FIG. 1. Schematic overview of the physical setup of the real-world phase in the guided EMA experiment, which took place inside a crowded university canteen over lunchtime.

experience. The monologues were derived from two fixed stories from the guide's personal life and were similar in content for each participant but not scripted.

During the 1-min assessments of the passive listening, communication, and active listening tasks, audio recordings (sampled at 48 kHz, 24 bit) were made with a custom-built, mobile microphone system, to obtain measurements of the broadband background levels as well as the guide's speech levels at the position of the participant. The system contained a cheek-mounted microphone and a reference microphone (both DPA 4066, DPA, Denmark), both connected to a portable audio recorder (Zoom H6 Handy Recorder, Zoom Corp, USA). The reference microphone was mounted vertically on a stand and placed at ear-height next to the participant, the same distance away from the guide. The cheek microphone was worn by the guide, who also operated the recording system. The system was calibrated before each experiment by recording the digital level measured in the reference microphone with a 1 kHz pure tone calibrator at a sound pressure level (SPL) of 94 dB (B&K type 4231, Brüel & Kjaer, Denmark). To account for varying pre-gain settings inside the recording device, the cheek microphone was calibrated in response to a brief, scripted message spoken by the guide while wearing the microphone inside an anechoic chamber at a distance of 1 m to the reference microphone. This procedure allowed for the derivation of a fixed scaling factor representing the broadband, free-field decay in speech level from the mouth of the guide to the reference microphone position at 1 m.

The microphone recordings were analyzed to derive broadband-level estimates of the background noise and the speech of the guide. The overall background noise level during the passive listening tasks was determined by calculating the average power in the reference microphone's recorded signal in dB SPL. To derive the guide's speech level at the participant's position during the communication and active listening tasks, the calibrated scaling factor was applied to the speech segments in the cheek microphone recording, which had been extracted using an adaptive, energy-based voice activity detector (VAD) (Kinnunen and Li, 2010). The background noise level was then computed from the reference microphone signal, using segments where the guide was not speaking (using the VAD derived from the synchronized cheek microphone signal). To obtain the background noise level during the communication task, the voice of the participant was also removed from the reference microphone signal using an additional energetic VAD. The details of the VADs used are described in Mansour *et al.* (2019a).

B. Laboratory assessment

Following the RW assessment, each participant repeated the passive and active listening tasks while seated

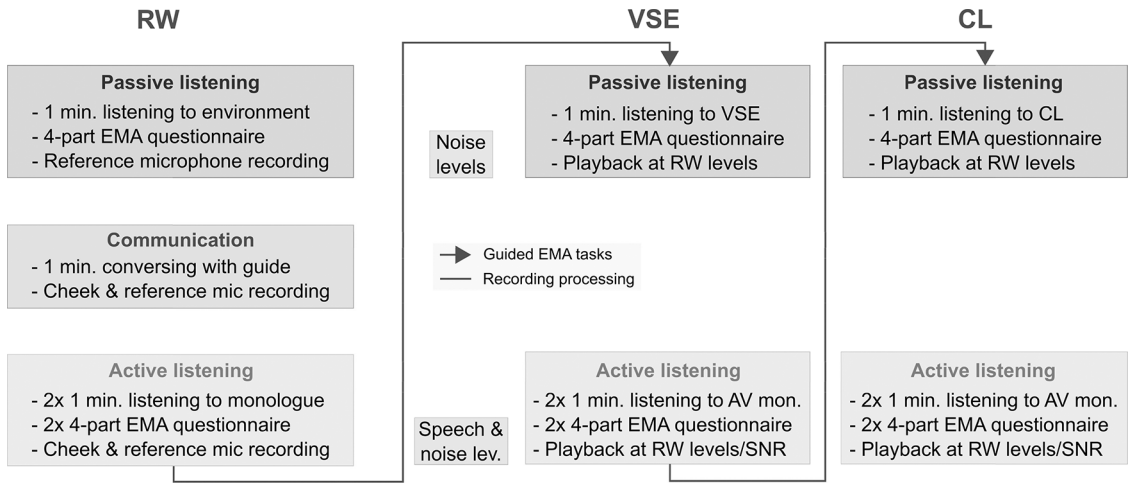

FIG. 2. Diagram depicting sequence of the RW, VSE and CL phases in the guided EMA experiment, as well as the passive listening, communication, and active listening stages within each phase.

inside an anechoically enclosed, 64-channel spherical loudspeaker array. This environment represented the fully spatialized VSE condition. The loudspeaker array, with a 2.4 m radius, used a spatial reproduction of a pre-recorded background noise signal to simulate the canteen environment. The background noise signal was recorded by a 32-channel spherical microphone array (em32 Eigenmike, MH Acoustics, USA), placed at head height inside the canteen, in the position where the participant was seated during the RW assessment. This 2-min-long recording was then encoded to a 4th-order Ambisonic signal and subsequently decoded to the geometry of the loudspeaker array for the VSE condition. In the CL condition, only two loudspeakers in the array were used, positioned at $\pm 90^\circ$ azimuth and 0° elevation, to present the background noise, mimicking a simple, audiology clinic setup. The two loudspeaker signals were derived from the same Ambisonic signal by decoding them to a binaural reproduction (Weisser and Buchholz, 2019) and subsequently applying a diffuse-field equalization step to account for the path between the loudspeakers and the ears of a head-and-torso simulator (B&K type 4128, Brüel & Kjær A/S, Denmark) located in the sweet spot. Both noise conditions were calibrated using the reference microphone, positioned in the acoustic sweet spot of the loudspeaker array.

The passive listening task was carried out by the participant in both the VSE and CL conditions, identical in structure to the corresponding tasks in the RW condition, as is summarized in the center and right columns of Fig. 2, respectively. To simulate the background noise, 1-min-long excerpts of the decoded noise signals were played back at the same participant-dependent broadband level as was measured in the real world. For the active listening task, the guide's monologue was simulated using pre-recorded, 1-min-long AV recordings spoken by the guide, derived from a personal story. For each participant, four monologues on four different topics were selected randomly out of a total of 16. The monologues were filmed inside an anechoic chamber on a neutral background using a digital camera (Sony a6000, Sony, Japan), and simultaneously recorded with the reference microphone, both at a distance of 1 m. The microphone recordings were calibrated and stored in both an unprocessed, single-channel format as well as a 64-channel, spatialized format obtained by convolving the single-channel signal with a spatialized room impulse response (RIR). This RIR was constructed by deriving the 64-channel Ambisonic loudpeaker signals from a spatial RIR recording, captured in quiet with the microphone array positioned at head height on the participant's chair inside the canteen. The resulting speech signals were then scaled to match the levels for the corresponding conditions in the real world and superimposed onto the background noise signal. The spatialized speech was used for the VSE condition, while the anechoic speech was presented from the frontal array loudspeaker (0° azimuth and elevation) in the CL condition. All speech signals were processed to provide 1 s of silence in the beginning and at the end.

The video recordings were played back on an 11-in. iPad (Apple, USA) positioned 1 m in front of the participant at eye level. The recordings were synchronized to the speech playback by using the video camera's own recorded speech signals to derive the delay with those of the reference microphone via cross correlation. Since the video signals were played back using a VLC media player, and not MATLAB (Mathworks, USA), the additional processing delay between both signal paths was taken into account by recording the video camera audio signal played back over the iPad at the same time as the loudspeaker array audio signal and using cross correlation to derive the constant offset between both signals.

C. EMA questionnaire design

Table I describes the EMA questionnaires for the passive and active listening stages, designed specifically for this experiment, detailing individual questions as well as their title in the smartphone app. The passive-stage questions focused on spatial sound perception (Q1), sound quality (Q2-Q3) and ability to relax (Q4), while the active-stage questions probed loudness perception of speech (Q5), listening effort (Q6), and quality of speech (Q7) as well as self-assessed speech understanding (Q8). These topics were chosen for their commonality in existing EMA research as well as in retrospective questionnaires like the SSQ. The response scale at the bottom of Table I indicates the possible responses on a 5-point Likert scale (Likert, 1932) for all questions except the final one (Q8), which was rated on a continuous scale from 0%–100%. Questions Q1–Q7 were

TABLE I. EMA questionnaires for the passive listening and the active listening stages and their respective title in the smartphone app, as well as the 5-point Likert response scale.

Passive listening stage		
No.	Title	Question
Q1	Difficulty to focus	Is it difficult for you to focus on specific sounds in this environment?
Q2	Pleasantness of sound	Does this environment sound pleasant to you?
Q3	Annoyance with sound	Are you annoyed with certain sounds in this environment?
Q4	Effort to relax	Is it effortful for you to relax in this environment?
Active listening stage		
Q5	Loudness of speech	How loudly did you feel the person you were talking to was speaking?
Q6	Listening effort	How effortful was it to listen to the person talking to you?
Q7	Naturalness of speech	How naturally did you think the person was talking to you?
Q8	Understanding of speech	How well did you understand what the person talking to you was saying?
Response scale		
Not at all… (1) - Not that… (2) - Somewhat… (3) - Very… (4) - Extremely… (5)		

https://doi.org/10.1121/10.0006568

phrased to fit the 5-point Likert response scale in order to provide a rigid and consistent structure to the questionnaires. The continuous answer scale of Q8 was chosen to allow the derivation of a self-assessed speech understanding score, similar to speech intelligibility (SI) scores produced by SI paradigms. Because of the known RW location, no questions were needed to establish the nature of a participant's surroundings. To reduce the variability in the RW environment between the assessment stages, the number of questions in both the passive and active listening questionnaires was limited to four, ensuring that the entire RW session could be completed within 30 min.

D. Participants

Fifteen participants with self-reported normal hearing carried out the guided EMA experiment. The participants were between 21 and 39 y, with a median age of 25 y, and all had English as their native language. The participants were recruited from the general public using a web advertisement and were financially compensated for their time. All regulations and guidelines were adhered to with regard to hygiene and social distancing, as brought about by the COVID-19 pandemic. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

III. RESULTS

A. Real-world noise and speech levels

Figure 3(a) shows the distributions of broadband noise SPLs (left), measured inside the canteen during the RW passive, communication, and active listening stages as well as the guide's speech SPLs (middle) and SNRs (right) during the communication and active listening stages. The left SPL ordinate indicates the noise data, while the right SNR ordinate represents the SNR data. The means (circles) and standard deviations (squares) of the distributions are indicated as well. The normality of each group was verified with the Anderson–Darling and Shapiro–Wilk tests. Mean noise levels were 65.8 dB SPL during the passive listening stage, 66.8 dB SPL during the communication stage and 66.4/65.6 dB SPL during the active listening stage at normal and challenging speech levels, respectively. As verified with a repeated-measures analysis-of-variance test (RANOVA), the noise distributions, though ranging from 62.5 dB SPL to 69.5 dB SPL, were not significantly different from each other $[F(3,42) = 1.15, p = 0.34]$. The mean speech level during the communication stage was 62.4 dB SPL, resulting in an SNR distribution with a mean of −4.5 dB. For the active listening stages at normal and challenging speech levels, the mean speech levels were 63.3 dB SPL and 53.6 dB SPL, respectively, yielding SNR distributions with mean values of −3.1 dB and −11.9 dB. Similarly, there was no significant difference between communication stage speech levels and normal speech levels in the active stage $[F(1,14) = 2.70, p = 0.13]$, nor between communication SNRs and normal, active-stage speech SNRs $[F(1,14) = 3.37, p = 0.09]$. These results confirm that despite natural day-to-day fluctuations, the background noise levels in the canteen stayed constant over the course of the RW assessment phase, which was intended to limit the intra-subject variability in the guided EMAs. In addition, the speech levels and SNRs remained constant between the communication task and the normal active listening task, implying that normal speech levels in the active listening task could be established using the communication task. For each of the participants, the maximum difference between the four background noise levels, as well as the maximum difference between the communication speech level and the normal active listening level, never exceeded 3 dB. The

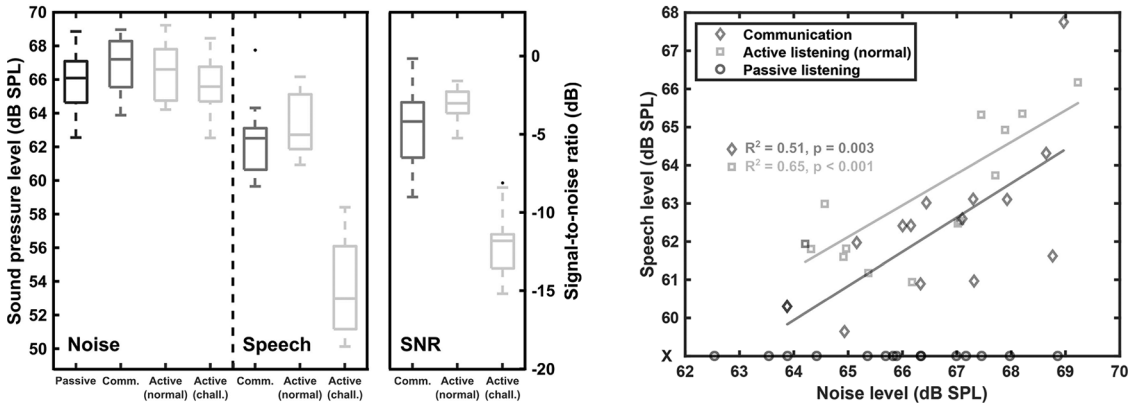

FIG. 3. Panel (a) Distributions of broadband noise SPLs (left), measured inside the canteen during the real-world passive, communication and active listening stages as well as the guide's speech SPLs (middle) and SNRs (right) during the communication and active listening stages. The noise and speech data use the left ordinate (in dB SPL), while the SNR data follow the right ordinate (in dB). The means (circles) and standard deviations (squares) of the distributions are indicated as well. Panel (b) Noise levels versus speech levels for the communication task and the active listening task at normal speech levels, along with least-squares fits to the data and its R^2 correlation value and significance p. The passive noise levels are shown on the bottom x axis and do not have a corresponding speech level (marked by X). Least-squares fits between corresponding speech and noise levels are shown with their R^2 correlation factor and goodness-of-fit p-value.

subjectively determined, challenging speech level for the normal-hearing participants was found to occur at a level of more than 10 dB below the background noise level, fluctuating more widely than the normal speech level.

Both the background noise levels [$F(14,42) = 11.1$, $p < 0.001$] and the guide's speech levels [$F(14,14) = 9.41$, $p < 0.001$] and SNRs [$F(14,14) = 9.41, p < 0.001$] did, however, vary significantly across different participants, despite the imposed time-of-day and location restrictions. Figure 3(b) shows the individual speech levels occurring during the communication task and the active listening task at normal speech levels as a function of their respective noise levels, as well as the noise levels during the passive listening task (shown on the bottom x axis, without a corresponding speech level as marked by X). Least-squares fits between corresponding speech and noise levels are shown with their R^2 correlation factor and goodness-of-fit p-value. The results highlight the varying background noise levels across participants during the passive listening tasks and indicate that the speech and noise levels were significantly positively correlated across a similar range during the communication task and the active listening task at normal speech levels. This correlation is in agreement with the established increase in speech effort as well as level with increasing background noise level (known as the Lombard effect) (Lombard, 1911) and explains the observed effects of participant in the statistical analysis. The Lombard slopes for speech levels during the communication task and active listening task were 0.88 dB/dB and 0.84 dB/dB, respectively. Particularly for the active listening task at normal speech levels, the least squares fit matches the data very well and is evident from the low variance observed in the SNR distribution of the active listening task at normal speech levels.

B. EMA responses

Figure 4 displays the participants' responses to the questions in the passive and active listening stages of the guided EMA. For each question, the number of responses for the RW, VSE, and CL conditions are given for each possible response on the 5-point Likert scale. The saturation of each response box corresponds to the relative frequency of the response. Each active listening stage question is divided into responses at normal (Norm.) speech levels and at challenging (Chall.) speech levels. Due to the non-parametric nature of the categorical output data, a Friedman test was applied, combined with Wilcoxon-rank *post hoc* tests with Bonferroni correction, to investigate differences between the conditions. The summary statistics are displayed underneath each question title and the significances of the *post hoc* results are indicated on the right side of each table.

Of the passive stage questions (Q1–Q4), only Q1 revealed significant differences between any of the conditions, specifically between the RW condition and both VSE and CL conditions. This implies that it was more difficult for participants to focus on specific sounds in the laboratory environments than in the real world (Q1). There were no significant differences between any of the conditions with

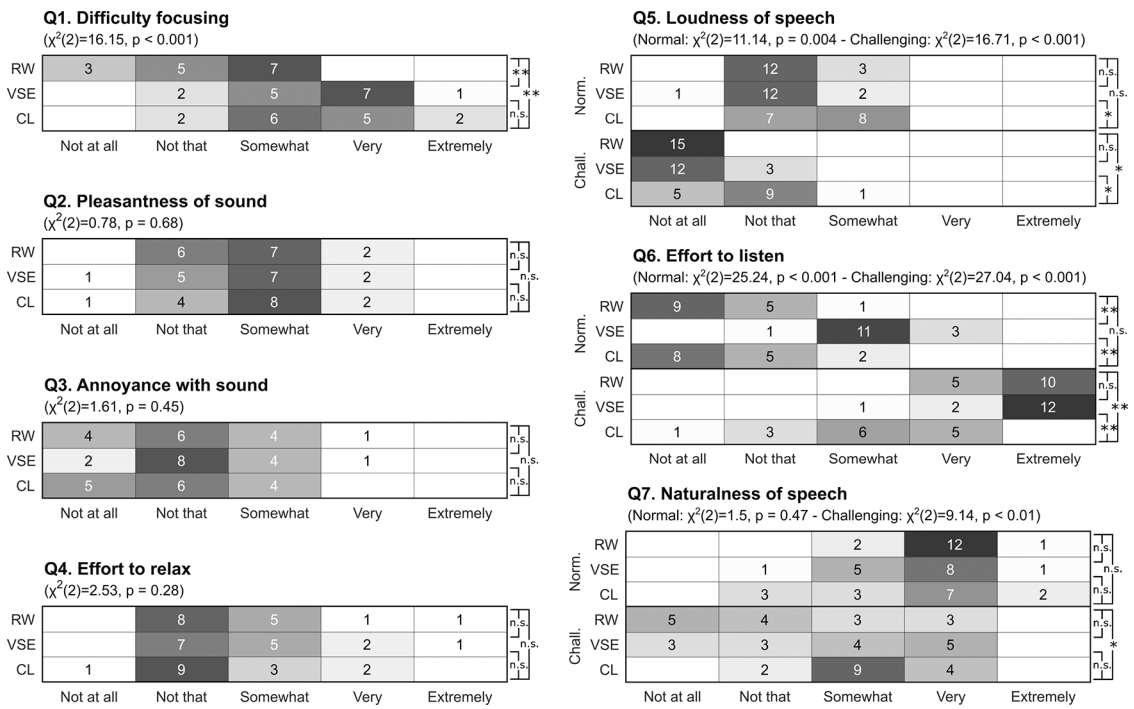

FIG. 4. EMA questionnaire responses for the passive listening (left) and active listening (right) stages. For each question, the number of responses for the real-world (RW), VSE, and clinic (CL) condition are given for each possible response on the 5-point Likert scale. The saturation of each response box corresponds to the relative frequency of the response. The summary statistics (Friedman test) are displayed underneath each question title and the significances of the *post hoc* results (Wilcoxon-rank test with Bonferroni correction) are indicated to the right of each table. Each active listening stage question is divided into responses at the normal speech level (Norm.) and at the challenging (Chall.) speech level.

https://doi.org/10.1121/10.0006568

regard to the experienced pleasantness of sound (Q2), the annoyance with sound (Q3) or the effort it took to relax (Q4). Participants generally agreed that the public lunch environment sounded mainly "somewhat pleasant" and "not that annoying", and was "not that effortful" to relax in. Except for Q1, these results support the absence of significant deviations in EMA results between the real world and the laboratory environments.

When actively listening to the guide's speech, its loudness was perceived as significantly higher in the CL condition than in the VSE condition, both at normal and challenging speech levels (Q5). At challenging speech levels, the RW condition was perceived significantly softer than the CL condition, but equally soft as the VSE condition. There were no significant differences between the RW and VSE conditions, which were perceived as mainly "not that loud" and "not at all loud" when judged at normal and challenging speech levels, respectively. These results indicate that while the VSE provided the expected consistency in loudness perception of speech to the real world, the clinic environment did not.

With regard to the participants' listening effort (Q6), the VSE condition was perceived as significantly more effortful than both the RW and CL conditions at normal speech levels, which were both perceived as mainly "not at all effortful". This difference disappeared at challenging speech levels, where the listening effort was now lower in the CL condition than in the RW and VSE conditions, where listening was perceived as mainly "extremely effortful". In contrast to what was expected, the listening effort at normal levels was thus higher in the VSE than in the real world, despite its equal perceived loudness. While the clinic environment reflected real-world listening effort at normal speech levels, it resulted in an "underestimation" of listening effort at challenging speech levels, unlike the VSE which was now similar to the real world. However, the clinic environment's similarity in listening effort to the real world at normal speech levels may have been caused by a flooring effect on the response scale.

No significant differences in the naturalness of the speech (Q7) were observed between the RW, VSE, and CL conditions at normal speech levels, perceived everywhere as mainly "very natural". At challenging speech levels, there was only a significant difference between the RW and CL conditions, whereby the RW condition was considered as mainly "not at all natural" compared to the CL condition being perceived as mainly "somewhat natural". As expected, the real-world naturalness of speech was thus preserved inside the laboratory environments at normal speech levels. At challenging speech levels, the naturalness of speech in the real world, as well as the VSE, was reduced, potentially due to the level of the speech stimulus being considered unnaturally low. This effect was partially mitigated in the CL condition due to the increased perceived loudness of speech (Q6).

Interestingly, there were no significant differences between participants for any of the passive listening questions and for any of the active listening questions at normal speech levels, despite the indicated fluctuations in speech and noise levels. All active listening questions at challenging speech levels contained a significant effect of participant. These observations support the guided EMA method's ability to produce, under normal listening circumstances, consistent assessments across participants. At challenging speech levels, the variability between participants increased, potentially due to the greater variance in SNRs and the absence of ceiling effects.

Figure 5 shows the distributions of percentage speech understanding for the final active listening question about self-assessed speech understanding (Q8) in each condition (RW, VSE, CL) at normal (Norm.) and challenging (Chall.) speech levels. A one-way repeated-measurement ANOVA (RANOVA) showed a significant effect of condition, both at normal and challenging speech levels. As verified by paired-samples t-tests, speech at normal levels was understood significantly less well in the VSE condition than in both the RW and CL conditions, which were not significantly different from each other. This difference disappeared at challenging speech levels, where the RW and VSE conditions were no longer significantly different even though the VSE condition remained somewhat more challenging than the RW condition. Here, speech understanding in the RW and VSE conditions was significantly lower than in the CL condition. Similar to Q6, the clinic environment seems to have been affected by a ceiling effect at normal speech levels, whereas the VSE reflected the real world most accurately at challenging speech levels.

C. Psychometric functions

Figure 6 shows psychometric functions of the participants' self-assessed speech understanding scores,

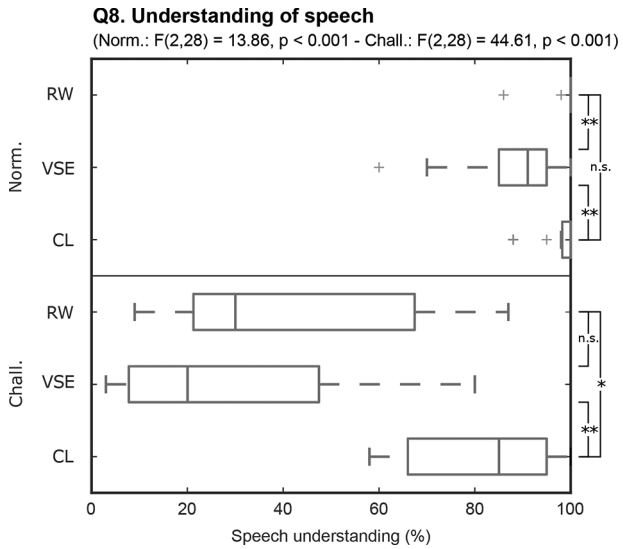

FIG. 5. EMA questionnaire responses for question 8 on self-assessed speech understanding. Panel A shows the ratings for the real-world (RW), VSE, and clinic (CL) condition, at the normal (Norm.) and challenging (Chall.) speech levels. The summary statistics (one-way repeated-measurement ANOVA) are displayed underneath the question title and the *post hoc* results (paired-samples t-test) are indicated to the right of the table.

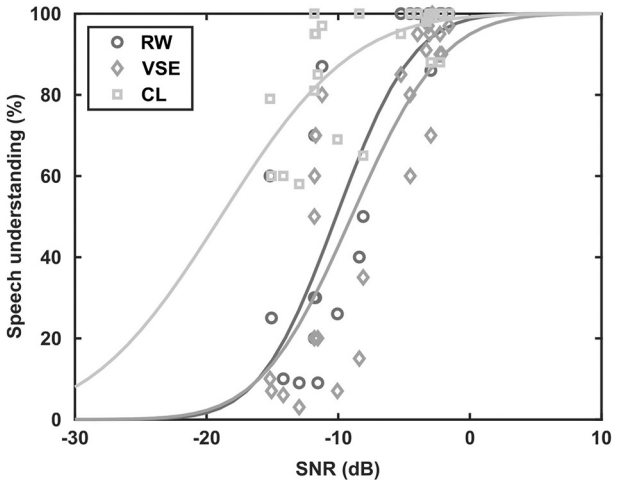

FIG. 6. Psychometric curves fitted to the participants' self-assessed speech understanding scores in question 8 as a function of the corresponding speech SNRs established in the real-world (RW), VSE, and clinic (CL) environments.

representing each score as a function of the corresponding speech SNR established in the real-world and laboratory environments (see Fig. 3). The psychometric functions for the RW and VSE conditions are very similar in their overall range and slope, with a 50% correct SNR of −10. 2 dB and −9 dB, respectively. The CL psychometric function has a shallower decay towards far more negative SNRs, with a 50% correct SNR at –19 dB. R^2 goodness-of-fit values for the psychometric functions are 0.69 for the RW condition, 0.60 for the VSE condition, and 0.53 for the CL condition. This indicates that self-rated speech understanding in the VSE resembled real-world values very closely, while the clinic environment resulted in both substantially increased speech understanding ratings as well as a poorer fit. The poorer quality of the clinic environment psychometric function is exacerbated by the absence of data points below 60% correct understanding.

IV. DISCUSSION

A. Real-world noise and speech levels

Due to the highly structured design of the guided EMA experiment and the presence of a guide, participants were able to carry out the EMAs without fail, indicating the full compliance of the participants in the proposed guided EMA method. The participant burden was reduced by the help of the guide and the limited assessments required by the participant. By using the same RW environment across all participants, selected for its importance, common occurrence and difficulty in people's lives, and limiting the experiment in time, the method also aimed to reduce inter- and intra-participant data variability. The observations from Fig. 3 showed that the background noise levels in the RW environment were consistent across the different assessment stages and that the guide's speech levels during the communication stage were similar to their normal speech levels during the active listening stage, as intended. Thus, the RW environment was acoustically stable in terms of sound levels over the course of the RW assessment stage, despite modest fluctuations across different participants caused by the changing distribution and number of interferers in the RW environment. Interestingly, the normal-speech-level SNRs were consistently negative around −4 dB, implying that conversational SNRs between normal-hearing interlocutors reached values below 0 dB even at noise levels below 70 dB SPL. Noise levels necessary to produce negative SNRs were reported to be over 5 dB higher in other studies (Pearsons *et al.*, 1977; Weisser and Buchholz, 2019). This may partially have been caused by the method with which the speech and noise levels were derived from the microphone recordings in the current study, which has been shown to result in lower, yet more accurate, SNR estimates (Mansour *et al.*, 2019a).

B. EMA responses

With regard to data variability and laboratory applicability, the passive stage questionnaire results (Q1–Q4) seem to have provided focused and consistent responses across participants (in favor of the hypothesis on the consistency between participants' real-world EMAs) and, with the exception of Q1, across environmental conditions. The highly similar perception of pleasantness of sound (Q2), annoyance with sound (Q3), and effort to relax (Q4) between the RW, VSE, and CL conditions further indicates that both laboratory environments could reproduce these sensations realistically (in favor of the hypothesis on the similarity between real-world and laboratory EMAs). The increased difficulty of focusing on specific sounds (Q1) in the laboratory environments was likely due to the absence of realistic visual stimuli, which are known to aid sound source localization (Shelton and Searle, 1980), as well as resulting from the limited spatial resolution the VSE could provide (Huisman *et al.*, 2020).

Similarly, the active stage assessments yielded consistent RW responses between participants at normal speech levels (in favor of the hypothesis on the consistency between participants' real-world EMAs), with the variance increasing somewhat at challenging levels. However, there were significant differences in the performance of the two laboratory environments relative to the real-world environment (in contrast to the hypothesis on the similarity between real-world and laboratory EMAs).

First, despite the presence of an AV speech stimulus, the VSE caused an increased listening effort and speech understanding difficulty compared to the real world. This was likely a consequence of the imperfect sound field reconstruction of the target speech by the 4th order Ambisonics system which has been shown to negatively affect speech intelligibility (Favrot and Buchholz, 2009). The absence of the differences in listening effort and speech understanding at challenging levels may have been caused by the overall perceived difficulty of the task. Nevertheless, the similarity between the psychometric functions of the VSE and the real

https://doi.org/10.1121/10.0006568

world (see Fig. 6) suggests that the VSE-based guided EMA task could discriminate self-rated speech understanding in a similar way to the real world at challenging speech levels. Second, the greater perceived loudness of speech (Q5), reduced listening effort (Q6) and shallower psychometric function resulting from the CL condition compared to the VSE and RW conditions suggests that the clinic environment allowed for an overall easier perception of speech in noise than the VSE and the real world.

The perceptual differences between the VSE and CL conditions might have originated from the acoustic differences between the single-loudspeaker anechoic CL speech source and the 64-loudspeaker reverberant VSE speech source. The absence of reverberation in the single-loudspeaker CL speech likely increased its intelligibility compared to the more reverberant VSE speech (Duquesnoy and Plomp, 1980). This is consistent with values of the speech transmission index (Steeneken and Houtgast, 1980), computed based on the VSE and CL speech stimuli, resulting in an average value of 0.98 for the CL stimulus compared to 0.89 for the VSE stimulus. This indicates that the CL speech is transmitted nearly perfectly, in contrast to the VSE speech. In addition, the spectral dissimilarities between the VSE and CL speech stimuli may have further contributed to the differences in perception. Figure 7 displays the difference, i.e., the SNR, between the long-term average speech spectra (LTAS) of the speech and the background noise in the VSE and CL conditions. All signals were binaurally recorded inside the loudspeaker array, averaged over both ears and normalized relative to their average power. The LTAS SNR for the CL condition was increased compared to that of the VSE condition, particularly in the region between 700 Hz and 1 kHz and anywhere above 1.5 kHz. Since the LTAS for the VSE and CL noise types was very similar (average power difference of less than 1 dB), the differences in LTAS SNR were primarily caused by the greater power contained in the CL speech compared to the VSE speech.

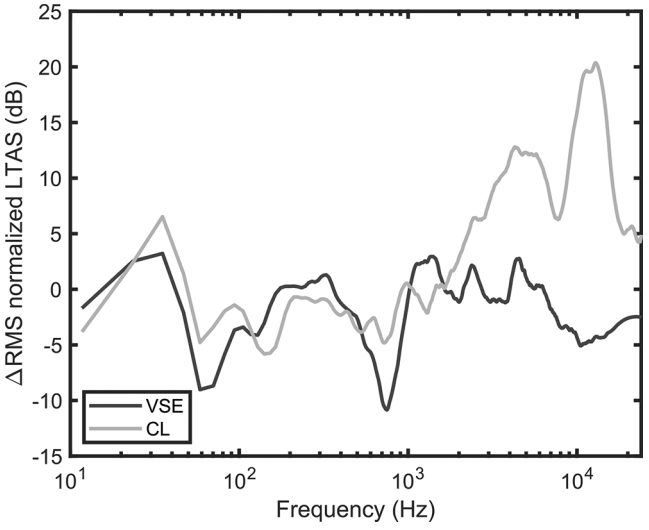

FIG. 7. Long-term average spectra of the VSE and CL noise and speech, binaurally recorded inside the loudspeaker array and averaged over both ears and normalized relative to their average power.

C. Limitations and outlook

The high naturalness with which speech was perceived by participants across conditions at normal speech levels showed that the guided EMA methodology could elicit natural listening experiences, in the real world as well as in the lab. Nevertheless, more efforts can still be made to improve the realism of the laboratory environments. Particularly the inclusion of more realistic visuals of the surroundings in addition to the target talker video would bolster the validity of EMA inside laboratory environments even further. The fact that participants always assessed the real-world environment first may have biased some of the EMAs due to the prior knowledge of what the environment was supposed to look and sound like. This was a necessary constraint, since the real-world speech and noise signals needed to be captured to inform the reproduction inside the laboratory environments. Finally, an important next step would be to apply the guided EMA methodology to hearing-impaired individuals, a participant group which was not included in the current study due to restrictions imposed by the COVID-19 pandemic. By evaluating hearing-impaired participants, unaided as well as aided by a hearing device, differences in subjective hearing ability between participants, as well as the effects of wearing a hearing device, could be captured. As such, the method of guided EMA could be further validated and potentially used to relate subjective hearing ability to objective measures of hearing loss and hearing device processing, both in the real world as well as in the lab.

V. CONCLUSION

This study explored a guided approach to EMA, which was designed to represent high feasibility, high participant compliance and low burden, as well as low inter- and intra-subject variability. The method, comprised of a passive listening, communication, and active listening task, was carried out with normal-hearing participants in a RW canteen environment and inside a VSE and a clinic laboratory environment, which were both acoustically matched with respect to the guide's speech and background noise levels. The real-world speech and background noise levels were shown to be similar across all participants, while the SNR necessary for natural communication was as low as −4 dB. The EMA results showed that the guided EMA methodology produced consistent passive listening EMAs within and across participants and environments. During active listening, the VSE generally resulted in EMAs most similar to the real-world environment, an observation which was supported by their highly similar psychometric functions. The clinic environment was perceived as less challenging, likely due to the increased intelligibility of its target speech source. The method of guided EMA may provide a new way of assessing subjective hearing ability in the real world, as well as in the lab, that can capture differences between participants and

relate them to objective acoustic or psychoacoustic outcome measures.

ACKNOWLEDGMENTS

The research was supported by Macquarie University and by the Centre for Applied Hearing Research (CAHR) at the Technical University of Denmark and by Widex A/S.

Ahrens, A., Marschall, M., and Dau, T. (**2017**). "Measuring speech intelligibility with speech and noise interferers in a loudspeaker-based virtual sound environment," J. Acoust. Soc. Am. **141**(5), 3510–3510.

Best, V., Keidser, G., Buchholz, J. M., and Freeston, K. (**2015**). "An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment," Int. J. Audiol. **54**(10), 682–690.

Duquesnoy, A., and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," J. Acoust. Soc. Am. **68**(2), 537–544.

Favrot, S., and Buchholz, J. M. (**2009**). "Validation of a loudspeaker-based room auralization system using speech intelligibility measures," in *Audio Engineering Society Convention 126* (Audio Engineering Society, Munich, Germany), p. 7763.

Galvez, G., Turbin, M. B., Thielman, E. J., Istvan, J. A., Andrews, J. A., and Henry, J. A. (**2012**). "Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users," Ear Hear. **33**(4), 497–507.

Gatehouse, S. (**1999**). "Glasgow hearing aid benefit profile: Derivation and validation of a client-centered outcome measure for hearing aid services," J. Am. Acad. Audiol. **10**, 80–103.

Gatehouse, S., and Noble, W. (**2004**). "The speech, spatial and qualities of hearing scale (ssq)," Int. J. Audiol. **43**(2), 85–99.

Henry, J. A., Galvez, G., Turbin, M. B., Thielman, E. J., McMillan, G. P., and Istvan, J. A. (**2012**). "Pilot study to evaluate ecological momentary assessment of tinnitus," Ear Hear. **33**(2), 179–290.

Huisman, T., Ahrens, A., and MacDonald, E. (**2020**). "Sound source localization with various ambisonics orders in virtual reality," J. Acoust. Soc. Am. **48**(4), 2786–2786.

Kinnunen, T., and Li, H. (**2010**). "An overview of text-independent speaker recognition: From features to supervectors," Speech Commun. **52**(1), 12–40.

Likert, R. (**1932**). "A technique for the measurement of attitudes," Arch. Psychol. **140**, 5–55.

Lombard, E. (**1911**). "Le signe de l'elevation de la voix (the sign of the elevation of the voice)," Ann. Mal. de L'Oreille et du Larynx (Ann. Dis. Ear Larynx) **37**, 101–119.

Lutman, M. E. (**1991**). "Hearing disability in the elderly," Acta Oto-Laryngol. **111**(sup476), 239–248.

Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (**2019a**). "A method for conversational signal-to-noise ratio estimation in real-world sound scenarios," J. Acoust. Soc. Am. **145**(3), 1873.

Mansour, N., Marschall, M., Westermann, A., May, T., and Dau, T. (**2019b**). "Speech intelligibility in a realistic virtual sound environment," in *23rd International Congress on Acoustics, Deutsche Gesellschaft Für Akustik eV* (September 9–13, 2019), pp. 7658–7665.

Moskowitz, D. S., and Young, S. N. (**2006**). "Ecological momentary assessment: What it is and why it is a method of the future in clinical psychopharmacology," J. Psychiatry Neurosci. **31**(1), 13–20.

Pearsons, K. S., Bennett, R. L., and Fidell, S. (**1977**). *Speech Levels in Various Noise Environments* (Office of Health and Ecological Effects, Office of Research and Development, US EPA, Washington D.C., USA).

Reis, H. T., and Judd, C. M. (**2000**). *Handbook of Research Methods in Social and Personality Psychology* (Cambridge University Press, London, United Kingdom).

Shelton, B., and Searle, C. (**1980**). "The influence of vision on the absolute identification of sound-source position," Percept. Psychophys. **28**(6), 589–596.

Shiffman, S., Stone, A. A., and Hufford, M. R. (**2008**). "Ecological momentary assessment," Annu. Rev. Clin. Psychol. **4**, 1–32.

Steeneken, H. J., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**(1), 318–326.

Timmer, B. H., Hickson, L., and Launer, S. (**2015**). "Adults with mild hearing impairment: Are we meeting the challenge?," Int. J. Audiol. **54**(11), 786–795.

Timmer, B. H., Hickson, L., and Launer, S. (**2017**). "Ecological momentary assessment: Feasibility, construct validity, and future applications," Am. J. Audiol. **26**(3S), 436–442.

Weisser, A., and Buchholz, J. M. (**2019**). "Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions," J. Acoust. Soc. Am. **145**(1), 349–360.

Westermann, A., and Buchholz, J. M. (**2015**). "The influence of informational masking in reverberant, multi-talker environments," J. Acoust. Soc. Am. **138**(2), 584–593.

Wolters, F., Smeds, K., Schmidt, E., Christensen, E. K., and Norup, C. (**2016**). "Common sound scenarios: A context-driven categorization of everyday sound environments for application in hearing-device research," J. Am. Acad. Audiol. **27**(7), 527–540.

Wu, Y.-H., Stangl, E., Zhang, X., and Bentler, R. A. (**2015**). "Construct validity of the ecological momentary assessment in audiology research," J. Am. Acad. Audiol. **26**(10), 872–884.