J Am Acad Audiol 26:872–884 (2015)

Construct Validity of the Ecological Momentary Assessment in Audiology Research

DOI: 10.3766/jaaa.15034

Yu-Hsiang Wu*
Elizabeth Stangl*
Xuyang Zhang*
Ruth A. Bentler*

Abstract

Background: Ecological momentary assessment (EMA) is a methodology involving repeated assessments/surveys to collect data describing respondents' current or very recent experiences and related contexts in their natural environments. The use of EMA in audiology research is growing.

Purpose: This study examined the construct validity (i.e., the degree to which a measurement reflects what it is intended to measure) of EMA in terms of measuring speech understanding and related listening context. Experiment 1 investigated the extent to which individuals can accurately report their speech recognition performance and characterize the listening context in controlled environments. Experiment 2 investigated whether the data aggregated across multiple EMA surveys conducted in uncontrolled, real-world environments would reveal a valid pattern that was consistent with the established relationships between speech understanding, hearing aid use, listening context, and lifestyle.

Research Design: This is an observational study.

Study Sample: Twelve and twenty-seven adults with hearing impairment participated in Experiments 1 and 2, respectively.

Data Collection and Analysis: In the laboratory testing of Experiment 1, participants estimated their speech recognition performance in settings wherein the signal-to-noise ratio was fixed or constantly varied across sentences. In the field testing the participants reported the listening context (e.g., noisiness level) of several semicontrolled real-world conversations. Their reports were compared to (1) the context described by normal-hearing observers and (2) the background noise level measured using a sound level meter. In Experiment 2, participants repeatedly reported the degree of speech understanding, hearing aid use, and listening context using paper-and-pencil journals in their natural environments for 1 week. They also carried noise dosimeters to measure the sound level. The associations between (1) speech understanding, hearing aid use, and listening context, (2) dosimeter sound level and self-reported noisiness level, and (3) dosimeter data and lifestyle quantified using the journals were examined.

Results: For Experiment 1, the reported and measured speech recognition scores were highly correlated across all test conditions ($r = 0.94$ to 0.97). The field testing results revealed that most listening context properties reported by the participants were highly consistent with those described by the observers (74–95% consistency), except for noisiness rating (58%). Nevertheless, higher noisiness rating was associated with higher background noise level. For Experiment 2, the EMA results revealed several associations: better speech understanding was associated with the use of hearing aids, front-located speech, and lower dosimeter sound level; higher noisiness rating was associated with higher dosimeter sound level; listeners with more diverse lifestyles tended to have higher dosimeter sound levels.

*Department of Communication Sciences and Disorders, The University of Iowa, Iowa City, IA 52242

Corresponding author: Yu-Hsiang Wu, Department of Communication Sciences and Disorders, The University of Iowa, Iowa City, IA 52242; E-mail: yu-hsiang-wu@uiowa.edu

This work was supported by a research grant from NIH/NIDCD R03DC012551, National Institute on Disability and Rehabilitation Research (NIDRR grant number H133E080006), and the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90RE5020-01-00). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS).

Portions of these data were presented at the Annual Meeting of the American Auditory Society in Scottsdale, AZ, March 2015.

The contents of this article do not necessarily represent the policy of NIDILRR, ACL, HHS, and the reader should not assume endorsement by the Federal Government.

Conclusions: Adults with hearing impairment were able to report their listening experiences, such as speech understanding, and characterize listening context in controlled environments with reasonable accuracy. The pattern of the data aggregated across multiple EMA surveys conducted in a wide range of uncontrolled real-world environment was consistent with the established knowledge in audiology. The two experiments suggested that, regarding speech understanding and related listening contexts, EMA reflects what it is intended to measure, supporting its construct validity in audiology research.

Key Words: Ecological momentary assessment, hearing aid, outcome

Abbreviations: CST = Connected Speech Test; EMA = ecological momentary assessment; HFA = high-frequency average of hearing loss; Leq = equivalent continuous sound pressure level; PDA = personal digital assistant; rau = rationalized arcsine units; SNR = signal-to-noise ratio

INTRODUCTION

For both researchers and audiologists, it is important to determine if a given intervention, such as a new hearing aid technology or fitting strategy, delivers greater benefit to listeners with hearing impairment than other interventions. The intervention benefits—or the outcomes—are often measured in a laboratory or clinic using methods such as speech recognition tests, or in the real world using retrospective self-reports such as questionnaires. Laboratory-type outcome measures have been widely used because they can assess outcome in controlled environments. On the other hand, retrospective self-reports have gained much attention in past decades because (a) the self-report nature is consistent with the trend toward a patient-driven health care system, (b) some domains of intervention outcome (e.g., satisfaction) cannot be assessed in laboratories, and (c) outcomes measured in the real world have better ecological validity (Cox, 2003).

Retrospective self-reports, however, have several disadvantages. First, they are subject to recall bias. Because retrospective self-reports are typically administered at least several weeks after intervention, such as hearing aid fitting, respondents have to recall and summarize their listening experiences across a long period of time. Empirical research has shown that long-term recall could be inaccurate and unreliable (Bradburn et al, 1987). For listeners who have lower cognitive abilities, accurately reporting real-world listening experiences in retrospective self-reports is even more difficult (Lunner, 2003).

Retrospective self-reports also suffer from poor contextual resolution. Specifically, many modern hearing enhancement technologies react and interact with the listening context, that is, the characteristics of listening activities, situations, and environments. If the listening context described in a questionnaire is not specific enough, the questionnaire might not be able to determine if a given technology is beneficial. For example, laboratory data have shown that, compared to omnidirectional microphones, the effect of directional microphone hearing aids on speech understanding could be positive, neutral, or even negative, depending on the locations of the talker and noises (Lee et al, 1998; Wu et al, 2013), signal-to-noise ratio (SNR) (Walden et al, 2005), reverberation level (Ricketts and Hornsby, 2003), and availability of visual cues (Wu and Bentler, 2010a, b). Because such detailed contextual information is not available in the questionnaire Abbreviated Profile of Hearing Aid Benefit (Cox and Alexander, 1995) (e.g., "When I am in a crowded grocery store talking with the cashier, I can follow the conversation."), it is unlikely that this inventory can detect the effect of directional technology in the real world (Ricketts et al, 2003).

Several techniques have been developed to overcome the disadvantages of retrospective self-reports. The ecological momentary assessment (EMA) is one of them. EMA, also known as experience sampling or ambulatory assessment, is a methodology involving repeated assessments/surveys to collect data describing respondents' current or very recent (i.e., momentary) experiences and related contexts in their natural (i.e., ecological) environments (Shiffman et al, 2008). In each assessment, experiences are recorded almost immediately; as a result, EMA is considered to be less affected by recall bias. Also, because detailed contextual information can be collected in each assessment, EMA has high contextual resolution.

EMA has been implemented using "low-tech" paper-and-pencil journals in previous hearing aid outcome research (Preminger and Cunningham, 2003; Walden et al, 2004; Cord et al, 2007; Wu and Bentler, 2010b, 2012). For example, to compare two hearing aid gain settings, Preminger and Cunningham (2003) asked participants to report the degree of listening difficulty and sound clarity of hearing aids in journals three times each day. For each journal entry, the participants also reported contextual information such as the setting of listening situation (e.g., restaurant/car) and noise level (quiet/low/high). Walden et al (2004) used the EMA methodology to explore the relationship between microphone preference (omnidirectional versus directional microphones) and listening environments. Hearing aid users were asked to report their preferred microphone modes in paper-and-pencil journals that used a check-box format. In addition to microphone preference, respondents also reported contextual information in terms of location of the listening activity (indoors/car/outdoors), location of the primary speech source (front/

side/back), location of background noise (front/side/back/all around), size of the indoor space (small/average/large), carpeting (presence/absence), and so on. The size of the room and carpeting were used to estimate the reverberation. The respondents were instructed to complete a survey whenever a major active listening situation (i.e., longer than a few minutes) occurred. In total, 1,599 journal surveys were completed by 17 hearing aid users. In a study designed to examine the effect of visual cues on directional microphone benefit, Wu and Bentler (2010b) expanded the survey used by Walden et al (2004) to collect more information. Hearing aid users were asked to report their degree of speech understanding using a 21-point scale. They also reported on contextual information in terms of the availability of visual cues (always/sometime/rarely) and relative loudness of noise compared to speech (much softer/somewhat softer/same/somewhat louder/much louder). Research participants were instructed to complete a survey whenever they encountered a predefined type of environment. In total, 1,367 surveys were completed by 24 hearing aid users.

A variant of low-tech EMA that has been used in hearing aid outcome research is a daily diary (Palmer et al, 2006; Bentler et al, 2008). For example, to evaluate the effectiveness of directional microphone hearing aids, Palmer et al (2006) asked research participants to complete diaries at the end of each day during the field trial. The participants used a scale ranging from "completely agree" to "completely disagree" to report if they agreed with statements such as "speech was more clear than usual today" and "noise was not as bothersome today." Although daily diaries do not ask respondents to record their immediate experiences, it could be considered a type of EMA due to the relatively short recall time frame compared to typical retrospective self-reports (Shiffman et al, 2008).

EMA can also be realized using "high-tech" portable computers (Galvez et al, 2012; Henry et al, 2012). For example, Galvez et al (2012) used personal digital assistants (PDAs) to characterize listening difficulty encountered by hearing aid users. Twenty-four hearing aid users were asked to carry PDAs for 2 weeks. The PDA prompted the participants through an audible alert to complete a survey four times per day. The questions of the survey were presented adaptively, depending on if respondents indicated experiencing any listening difficulties since the last survey. In total, 991 assessments were completed. Because the participants showed high compliance (77% response rate to the PDA alarm) and reported positive feedback, the study by Galvez et al (2012) supported the feasibility of computerized EMA. Due to the recent advancement of smartphone technology, applications/software that allow researchers to implement the EMA methodology using smartphones in outcome research has been developed (Hasan et al, 2013). The data seem to support the feasibility of using smartphone-based EMA in hearing aid outcome research (Hasan et al, 2014).

Although the use of EMA in audiology research is growing and its validity has been confirmed in other disciplines (Hektner et al, 2007; Shiffman et al, 2008), evidence supporting the construct validity, which is the extent to which a measurement reflects what it is intended to measure (Cronbach and Meehl, 1955), of EMA in audiology research is scarce. For example, EMA has been used to measure individuals' listening experiences such as the degree of speech understanding (Wu and Bentler, 2010b). For EMA to have high construct validity, respondents need to accurately estimate and report their degree of speech understanding in each assessment or survey. Although literature has shown that adults with hearing impairment preserve the ability to rate speech recognition performance, most of the previous research was conducted in laboratory environments wherein the test condition was fairly static (Cox et al, 1991; Cienkowski and Speaks, 2000; Wu and Bentler, 2010a). Because real-world environments can change quickly from moment to moment, it is unknown if the degree of speech understanding reported in EMA surveys would approximate what respondents actually experience in the real world.

To achieve high construct validity, EMA also requires respondents to accurately describe the characteristics of different listening contexts. Some contextual properties are more static and easier to be recognized (e.g., indoor versus outdoor location). However, reporting contextual characteristics that can change substantially from time to time (e.g., location of the primary talker) is more difficult. Therefore, it is unknown to what extent the listening context data collected in EMA surveys reflect what actually happens in the real world.

Finally, because it is impossible to strictly control real-world conditions and environments, EMA data are generally noisy. To derive a clear pattern of human experiences and behaviors, EMA relies on repeated assessments and data aggregation. If EMA reflects what it is intended to measure, the pattern of the data aggregated across EMA's multiple assessments should be consistent with established knowledge or theories. For example, it is well established that speech understanding decreases as noise increases. If EMA is a valid measure, aggregated EMA data should reveal an association between poorer speech understanding and higher noisiness rating. In the study by Walden et al (2004) the aggregated EMA data indicated that the directional mode was preferred over the omnidirectional mode when background noise was present and the speech source was located in front of and near the listener. Because this finding was consistent with the theoretical acoustic effect of directional microphones, the construct validity of EMA was somewhat supported. However, the purpose of Walden et al (2004) was to explore the unknown relationship between microphone mode preference and real-world environment. No study has been conducted to verify EMA's

construct validity in audiology research by examining the relationship between EMA data and established knowledge or theories.

The purpose of the two experiments presented in this article was to systematically examine the construct validity of EMA in terms of measuring speech understanding and related listening context. At the "micro level," Experiment 1 investigated if in a given assessment adults with hearing impairment could accurately (a) rate their speech recognition performance in a more dynamic laboratory setting and (b) characterize the listening context of semicontrolled real-world environments. At the "macro level," Experiment 2 investigated if the pattern of the real-world data aggregated across repeated EMA assessments would be consistent with established knowledge regarding the relationships between speech understanding, hearing loss, hearing aid use, listening context, and lifestyle.

EXPERIMENT 1

Several previous studies have shown that adult listeners can estimate their speech recognition performance in laboratory settings (Cox et al, 1991; Cienkowski and Speaks, 2000; Wu and Bentler, 2010a). However, these studies typically presented sentences at fixed SNRs and asked listeners to report their performance after listening to few sentences (ranging from 1 to 20 sentences) in a very short time frame. In contrast, in EMA surveys, respondents often have to estimate their speech understanding across a longer time frame (e.g., ≥10 min) in environments wherein the SNR (and thus speech intelligibility) changes quickly from moment to moment. To obtain an insight into the extent to which listeners can accurately report their speech understanding in EMA surveys, the first purpose of Experiment 1 was to investigate the relationship between reported and measured speech recognition performance in laboratory settings wherein the SNR was varied over a longer period of time.

The second purpose of Experiment 1 was to investigate if listeners could accurately report listening context properties in EMA surveys. Participants were asked to engage in conversations with two observers in various real-world environments. Both participants and observers described the listening context using paper-and-pencil journals after each conversation. The journal data were then compared to determine the consistency between participants and observers.

Methods

Participants

Twelve adults (six males and six females) participated in the experiment. Participants had to (a) have bilateral downward-sloping sensorineural hearing loss; (b) have a hearing threshold symmetry within 15 dB for all test frequencies; and (c) be able to understand the directions of experiments and conduct experimental tasks. Participants' ages ranged from 27 to 79 yr with a mean of 64.8 yr (SD = 17.5). The mean pure-tone thresholds are shown in Figure 1. All participants were experienced hearing aid users. A participant was considered an experienced user if he or she had used hearing aids >4 h per day in the previous year and kept using hearing aids during the study.

Laboratory Tests

To determine if the participants could accurately estimate their speech recognition performance, the Connected Speech Test (CST; Cox et al, 1987) was used. This sentence recognition test was chosen because it was designed to simulate everyday conversations in which speech contextual cues are usually available. The CST sentences are from a collection of passages about common topics. Each CST passage consists of nine or ten sentences.

Three conditions were created for the experiment. The first was the standard condition wherein the CST sentences were presented at fixed SNRs as the previous research. To eliminate the floor and ceiling effect, the test SNRs were adjusted for each individual listener. Specifically, before the formal testing, 20 CST sentences were presented to measure the SNR-50, at which the listener could understand 50% of speech, using adaptive SNR procedures. The multitalker babble of the CST was fixed at 60 dBA. The speech level was adjusted depending on the listener's responses using the one-down, one-up adaptive procedure in 2-dB steps. The correct response of each sentence was based on the repetition of the whole sentence, with minor exceptions such as "a" and "the." The presentation SNR averaged

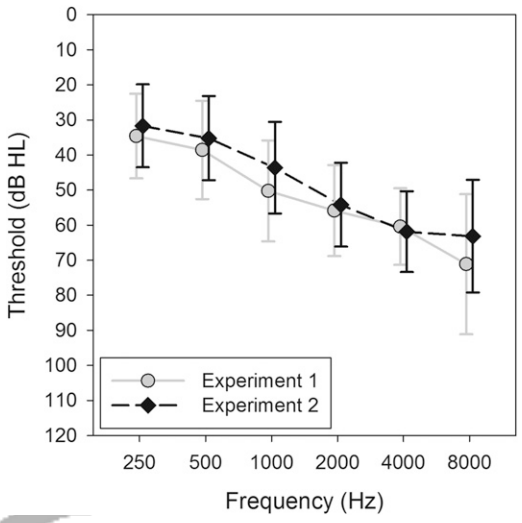

Figure 1. Mean hearing thresholds for participants in Experiments 1 and 2. Error bars indicate 1 standard deviation.

across sentences 5–20 defined SNR-50. Relative to an individual's SNR-50, three SNRs, −6, 0, +6 dB, were created with the babble level fixed at 60 dBA. For each SNR, a pair of CST passages (19–20 sentences) was presented. After listening to each sentence, the participants' task was to repeat as much of each sentence that they heard as possible. Performance was scored based on the number of key words correctly repeated out of the key words presented. After listening to a pair of CST passages, the participants reported their performance using a 21-point scale, ranging from understanding nothing (0%) to everything (100%) with the scale marked in 5% steps. The order of the three SNRs was randomized across participants.

The second test condition was the roving condition, which was identical to the standard condition except that the SNR roved from sentence to sentence. For each of the three SNRs (−6, 0, +6 dB relative to SNR-50), the sentence level was randomly altered by −2, −1, 0, 1, or 2 dB from the nominal SNR. One pair of CST passages was used in each nominal SNR. The participants' tasks were identical to the standard condition.

The third test condition was the long roving condition, which was similar to, but longer than, the roving condition. For each of the three SNR (−6, 0, +6 dB relative to SNR-50), three SNR blocks were created: −3, 0, and +3 dB relative to the nominal SNR. For example, the three blocks of the −6 dB SNR were −9, −6, and −3 dB SNRs. Within each block, the sentence level randomly roved by −2, −1, 0, 1, or 2 dB from the nominal SNR of that block. One pair of CST passages was used in each SNR block. After listening and repeating sentences for three SNR blocks (three pairs of CST passages; ∼ 60 sentences), listeners were asked to estimate their overall performance. The order of the three SNR blocks was randomized.

The test was administered in a laboratory space with low reverberation (reverberation time = 0.21 sec). The speech signals were generated by a computer with a Motu Ultralite-mk3 Hybrid sound interface (MOTU Inc., Cambridge, MA), routed via a GSI 61 audiometer (Grason-Stadler, Eden Prairie, MN), an 8-channel Alesis DEQ830 digital equalizer (Alesis, Cumberland, RI), and an ADCOM GFA5002 amplifier (ADCOM, Marlboro, NJ), and then presented from a Tannoy i5 AW loudspeaker (Tannoy Ltd., Coatbridge, Scotland) located at the listener's eye level at 0° azimuth. Uncorrelated CST babble was generated by another computer with a Focusrite Saffire multichannel sound interface, routed via the DEQ830 equalizer and GFA5002 amplifiers, and presented from eight Tannoy i5 AW loudspeakers located at 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315° azimuth. The distance between the listener and loudspeakers was 1.2 m.

Field Tests

To determine if listeners could correctly characterize listening context, the participant and two normal-hearing research assistants who served as observers moved to ten different locations and had conversations. The observers were trained to create natural conversations in various speech-listener configurations (e.g., face-to-face or side-by-side). The locations were selected so that the variation of acoustic properties of the environment could be maximized (e.g., from quiet to very noisy). The locations included a clinic waiting room, a coffee shop, and walking along outdoor streets. After having a conversation in each location for several minutes, the participant conducted a survey describing the context using a paper-and-pencil journal. The journal used a check-box format to characterize the listening context in terms of conversation location (outdoor traffic/outdoor nontraffic/indoor, ≤10 people/indoor, ≥11 people), primary talker location (front/others), noisiness level (quiet/somewhat noisy/noisy/very noisy), noise location of noisy environments (front/rear/side/all around), size of indoor space (small/average/large; compared to average living room), and carpeting of indoor space (yes/no). If the context (e.g., primary talker location) changed over time during the conversation, the participant was asked to select the one that occurred most of the time. The survey was adapted from Walden et al (2004) and Wu and Bentler (2010b).

To understand if the context was correctly characterized, after each conversation the two observers also answered the same survey questions in their own journals from the viewpoint of the participant. The participant and the observers were blinded to one another's answers. To examine if the self-reported noisiness would reflect the background noise level of the environment, the observers used a Larson-Davis System 824 sound level meter (Larson Davis Inc., Depew, NY) to measure the noise level. No conversation was conducted during the noise level measurement.

Procedures

The study was approved by the Institutional Review Board of the University of Iowa. After agreeing to participate in the study and signing the consent form, participants' pure-tone thresholds were measured. If participants met the inclusion criteria, laboratory testing was then administered, followed by the field testing. Before laboratory testing, the participant's SNR-50 was measured and a practice session was held to familiarize participants with the speech recognition and performance estimation tasks. In the formal testing, the order of the standard, roving, and long roving conditions was randomized. Before the field testing, a training session was given to ensure that the participants understood the survey questions.

Note that during all of the testing the participants used their own hearing aids. The hearing aids differed somewhat from each other but were all potentially appropriate for the participants' hearing loss. No verification measures were conducted and hearing aid features

were not logged in this study. Each hearing aid was worn at a volume control setting and program/memory selected by the participants. Also note that in the field testing the conversation location, talker–listener configuration and distance, and the background noise level were not controlled. Although the hearing aids would have an effect on the speech recognition in the laboratory testing and could modify the perception of sounds in the field testing, and although the listening context varied within and between participants, differences among hearing aids and listening contextual properties were not of interest in this experiment; the main focus of the experiment was the relationship between reported and measured CST scores (laboratory testing) and the consistency in survey results between the participant and the observers (field testing).

Results

Speech Recognition

Before analysis, the measured and reported CST scores were transformed into rationalized arcsine units (rau) to homogenize the variance (Studebaker, 1985). Figure 2 illustrates the relationship between the reported and measured scores in each of the standard, roving, and long roving conditions. In the standard condition, data are well described by the diagonal line, suggesting that reported and measured scores were very close. The linear correlation coefficient between reported and measured scores was 0.97 ($p < 0.001$). On the other hand, even though the data for the roving and long roving conditions are more dispersed, the correlations between reported and measured scores remained high (for both conditions: $r = 0.94$, $p < 0.001$).

To determine whether there were systematic differences between reported and measured CST scores, a repeated measures analysis of variance was conducted to examine the effect of score type (reported/measured), test condition (standard/roving/long roving), and SNR (−6/0/+6 dB) on CST scores. Results revealed a significant difference between the two types of score [$F_{(1,11)} = 7.12, p = 0.02$], with the mean measured score (55.6 rau) higher than the reported score (51.6 rau). The results further indicated that the main effect of SNR was significant [$F_{(2,22)} = 129.5$, $p < 0.001$]. The test condition main effect and all interactions were not significant.

Listening Context

The answers to survey questions regarding conversation location, room size, and carpeting were first used to derive the degree of reverberation (low versus high). Specifically, outdoors were assumed to have low reverberation. Indoor, carpeted spaces that were equal in size or smaller than an average living room were considered to have low reverberation. The remaining indoor locations were assumed to have high reverberation (Walden et al, 2004).

For each survey question and the degree of reverberation in each location, the results from the participant and the two observers were compared. The percent consistency for each participant was then calculated by dividing the number of consistent surveys by the number of total surveys. Figure 3 shows the mean percent consistency across all participants. The consistency between the two observers is also shown. Star symbols in the figure indicate the chance level of consistency (i.e., the percent consistency if participants and observers randomly chose the answer). The participants and the observers' answers were highly consistent (92–95%) in terms of conversation location, carpeting, and reverberation. The consistency was poorest for noisiness rating (58%). Seven one-sample t tests were conducted, one for each of the context properties shown in Figure 3, with the Bonferroni correction to examine if the consistency

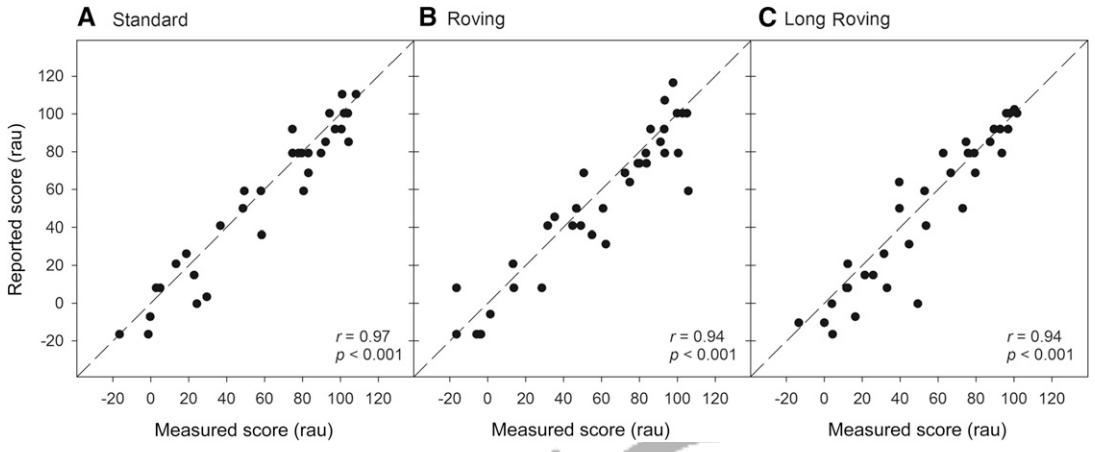

Figure 2. Reported speech recognition score as a function of measured score in the standard (A), roving (B), and long roving (C) conditions. Dashed diagonal lines represent perfect match.

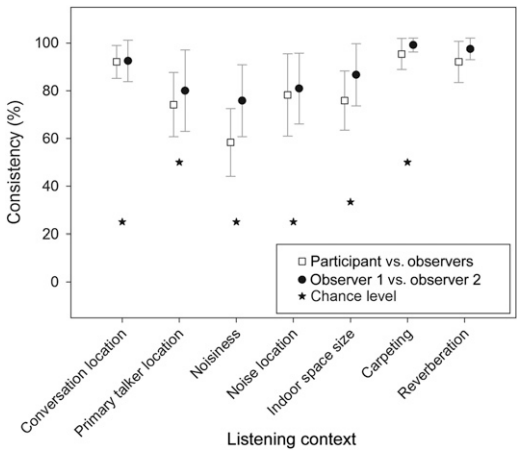

Figure 3. Consistency of reported listening context between research participants and observers and between the two observers. Error bars indicate 1 standard deviation.

between participants and observers was above the chance level. The results indicated that this is the case.

Figure 4 shows a box plot of the background noise level measured using a sound level meter as a function of noisiness rating reported by the participants. In general, noisiness rating increased monotonically as noise level increased. To determine if the trend shown in Figure 4 was statistically significant, ordinal logistic regression analysis was performed. This analysis used noise level in a repeated measure manner to predict the cumulative odds ratio of the probability of choosing a given noisiness rating (e.g., noisy) and the ratings that had higher noisiness levels (e.g., very noisy) to the probability of choosing the ratings with lower noisiness levels (e.g., quiet and somewhat noisy). Because hearing loss might affect the perception of noisiness, the effect of high-frequency hearing loss average (HFA; threshold averaged across 1, 2, and 4 kHz) was controlled for in the analysis. The result indicated that the effect of noise

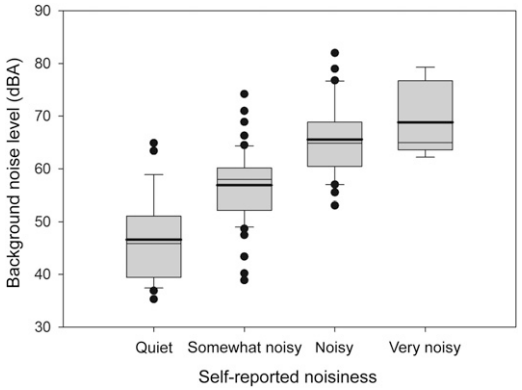

Figure 4. Box plot of background noise level as a function of self-reported noisiness. The boundaries of the box represent the 25th and 75th percentile. The thinner and thicker lines within the box mark the median and mean, respectively. Error bars indicate the 10th and 90th percentiles. Solid circles are outlying data points.

level was significant ($\chi_1^2 = 6.64$, $p = 0.01$), suggesting that the participants tended to report higher noisiness ratings in environments wherein the background noise level was higher. The effect of HFA was not significant.

Discussion

Although the reported CST score (51.6 rau) was systematically lower than the measured score (55.6 rau), the high correlations between the two types of scores indicated that listeners could estimate their relative speech recognition performance. The significant correlations across all three test conditions (Figure 2) further suggested that the estimations of performance were accurate not only in static environments with fixed SNRs, but also in more dynamic situations wherein the SNR was constantly changing for a longer period of time. Based on these results, it is likely that the degree of speech understanding reported in EMA surveys is accurate.

For most listening context properties investigated in the field test, the participants and the observers were generally consistent, indicating that participants could characterize the listening context. However, the consistency of noisiness rating was lower (58%; Figure 3). The low consistency is not too surprising because the perception of noisiness was more subjective than the other contextual properties. This can be illustrated by the low consistency between the two observers (76%; Figure 3). Furthermore, the participants and the observers could have different perceptions of noisiness because the former had hearing loss and used hearing aids while the latter did not. Hearing aid features such as compression, noise reduction algorithms, and directional microphones might alter the perception of noisiness (Wu and Stangl, 2013).

To further examine the pattern of noisiness inconsistency, the percentages of surveys in which the participants reported higher (noisier), same, or lower (quieter) noisiness levels than the observers were calculated in each of the four background noise level categories: <50, 50–60, 60–70, and >70 dBA. The noise level was measured using the sound level meter. The results (Figure 5) indicated that the consistency decreased as background noise level increased (the light gray bars). The data shown in Figure 5 also seem to support the effect of hearing aid features on the inconsistency in noisiness ratings between the participants and normal-hearing observers. Specifically, because compression algorithms make soft sounds louder, the participants were more likely to report higher noisiness levels than the observers in quieter environments (<60 dBA). On the other hand, because compression, noise reduction algorithms, and directional microphones (if activated automatically in noisier environments) make loud sounds/noises softer, the participants tended to report

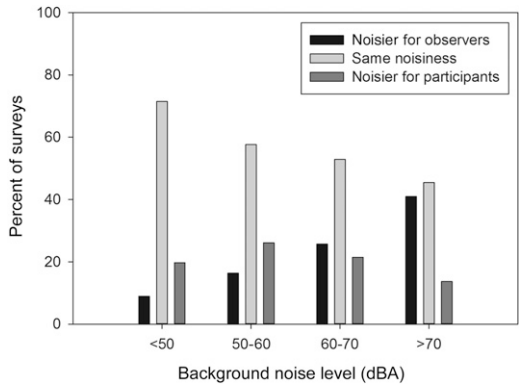

Figure 5. Percentages of surveys in which the participants reported higher, same, or lower noisiness levels than the observers as a function of background noise level.

lower noisiness levels in noisier environments (>60 dBA). If features' real-time statuses during the field testing were available, the effect of these features might be controlled for in statistical analyses and the consistency between participants and observers might increase.

Despite the lower consistency, noisiness ratings increased monotonically as the background noise level increased (Figure 4). Therefore, the results of the experiment suggested that although the noisiness ratings reported by the participants were not highly consistent with those rated by normal-hearing observers, the participants were able to estimate the noisiness level.

The results of Experiment 1 indicated that listeners with hearing impairment could estimate the relative degree of speech understanding and describe listening context with reasonable accuracy. These results, however, were unable to fully support the construct validity of EMA because real-world listening situations are often more complicated and dynamic than the laboratory settings and the semicontrolled conversations used in Experiment 1. Therefore, Experiment 2 was conducted to examine EMA's construct validity at the macro level.

EXPERIMENT 2

In this experiment, participants were asked to repeatedly report their degree of speech understanding and describe the listening context using EMA journals in their natural environments for 1 week. During that week, the participants also carried noise dosimeters to measure the sound level of the environment. This experiment was part of a larger study and portion of the larger study has been reported in Wu and Bentler (2012).

The rationale of this experiment was that, for EMA to have high construct validity, the results generated by this methodology should be consistent with the established knowledge or theories in audiology. Based on this rationale, three hypotheses were formulated. First, it is well established that audibility (Humes, 2002) and visual cues (Sumby and Pollack, 1954) play important roles in speech recognition. Therefore, it was hypothesized that, when aggregating across multiple EMA surveys completed in various listening situations, better speech understanding would be (a) associated with situations wherein the listener was using hearing aids (better audibility) and the primary talker was in front of the listener (visual cues might be available) and (b) negatively associated with the degree of hearing loss (poorer audibility). The second hypothesis involves the relationship between self-reported EMA data and dosimeter data. Specifically, although noise dosimeters do not directly measure SNR, the overall sound level collected by dosimeters can estimate SNR because of the high correlation between them (Pearsons et al, 1976; Banerjee, 2011). Therefore, it is hypothesized that better speech understanding and lower noisiness ratings (i.e., quieter) reported in EMA surveys would be associated with lower overall sound levels measured using the dosimeters. Third, research using retrospective self-reports has shown that adults with less active or less diverse lifestyles tend to experience more quiet environments (Wu and Bentler, 2012). The variation in environmental sound level is also smaller for these individuals (Gatehouse et al, 2006). Therefore, it was hypothesized that more active/diverse lifestyles derived using EMA data would be associated with higher overall and more varied sound level collected by the dosimeters.

Methods

Participants

Twenty-seven adults (7 males and 20 females) were recruited from the community and served as participants. The inclusion criteria were identical to Experiment 1. Participants' ages ranged from 40 to 88 yr with a mean of 66.3 yr (SD = 11). Twenty of the participants were experienced hearing aid users. The mean pure-tone thresholds are shown in Figure 1.

EMA Journal

The participants used paper-and-pencil journals to report their listening experiences and describe the listening contexts that they encountered in their everyday lives for a week. During the week, whenever the participants had a listening condition >10 min, they described the auditory activity and acoustic environment of that condition in the journal. The journal used a check-box format and provided six listening activity categories. Among them, three categories involved conversations (small group/large group/phone), two categories involved speech listening (live speech/media), and one category for

not actively listening to speech. The journal provided five environmental categories, including two outdoor (traffic/nontraffic) and three indoor (home/nonhome/crowd of people) categories. Combining 6 activities and 5 environmental categories, the journal provided 30 different listening events. In each survey, the participants were allowed to select only one activity and one environmental category. If they were performing more than one activity in a given listening condition (e.g., talking to friends while watching TV), they selected the activity that occurred most of the time.

If the listening event involved conversation or speech listening, the participants were further asked to report the degree of speech understanding using a 21-point scale ranging from understanding nothing (0%) to everything (100%) with the scale marked in 5% steps. The participants also reported if they were using hearing aids in that listening event (yes/no) and characterized the listening context in terms of the primary talker location (front/others) and noisiness level (quiet/somewhat noisy/noisy/very noisy). Finally, they recorded the starting and ending times of the event. The participants were asked to complete the survey immediately following the listening event.

Noise Dosimeter

The participants were asked to carry Larsen-Davis Spark 703 dosimeters during the week that they conducted EMA surveys. The Spark 703 dosimeter measured the A-weighted equivalent continuous sound pressure level (Leq) every 5 sec and logged the level data along with the time information to its internal memory. The Leq measurement range was set to span from 43 to 113 dBA. The dosimeter was programmed to start measuring and logging Leq data automatically each morning and to switch off each night. The on and off times were set in accordance with participants' daily schedule.

The dosimeter was placed in a 22 × 17 × 7-cm carrying bag with the microphone clipped to its outside. The length of the shoulder strap was adjusted so that the bag sat at waist level when carried on the participants' shoulders.

Procedures

After the participants completed the consent procedure, pure-tone thresholds were measured. If participants met the inclusion criteria of the study, a training session was given to ensure that they understood how a dosimeter works, how to carry the bag, and how and when to complete a survey. The dosimeter was then programmed and the internal clock of the dosimeter was synchronized to the participants' watches or cell phone clocks. During the next 7 days, participants were instructed to carry the dosimeter/bag, conduct surveys, and maintain their regular daily activities and schedules. Hearing aid users were encouraged keep using their hearing instruments as usual. Journals were printed in small notebook form, so that they could easily fit in the dosimeter bag. The participants were asked to carry the bag on their shoulders whenever possible. However, they were allowed to place the bag somewhere close to them (e.g., on a desk) given that they stayed within a 1-m radius of the bag. The participants were also encouraged to complete as many surveys as possible. One week later, participants returned to the laboratory to download the dosimeter data and turned in the journals.

Results

The data collected from dosimeters and journals were prepared before analysis. Surveys in which the participants indicated that they forgot to carry the dosimeter (n = 28) were eliminated. Using the time information recorded in each survey, the dosimeter 5-sec Leq data of a given listening event was extracted to calculate the Leq of the entire event. The survey and dosimeter data were then inspected again. The surveys in which the participants were very unlikely to have carried the dosimeter (e.g., a survey indicating "very noisy" while the event Leq was close to the lower measurement limit of the dosimeter) were eliminated (n = 16). The remaining data, which consisted of a total of 1,267 surveys covering 2,032 h of dosimeter recordings, were subjected to analysis.

Speech Understanding

Among the 1,267 surveys, 667 surveys involved conversations or speech listening. Figure 6 shows the box plots of speech understanding as a function of HFA, hearing aid use, primary talker location (Figure 6A), and event Leq (Figure 6B). To more efficiently present the data in the figures, the continuous variable HFA was categorized into two groups (better or worse than the average) and event Leq was divided into five categories (50–80 dBA with 10-dB steps). To determine how hearing loss (i.e., HFA; continuous variable), hearing aid use (categorical variable), primary talker location (categorical variable), and event Leq (continuous variable) would affect speech understanding (continuous variable), a mixed model that allowed errors of the same participant to be correlated was conducted. The results (Model A in Table 1) indicated that better speech understanding was significantly associated with lower HFA, the use of hearing aids, front-located speech, and lower event Leq.

Note that although better speech understanding was associated with lower event Leq, Figure 6B indicated that the lowest event Leq category (<50 dBA) did not generate the highest level of speech understanding.

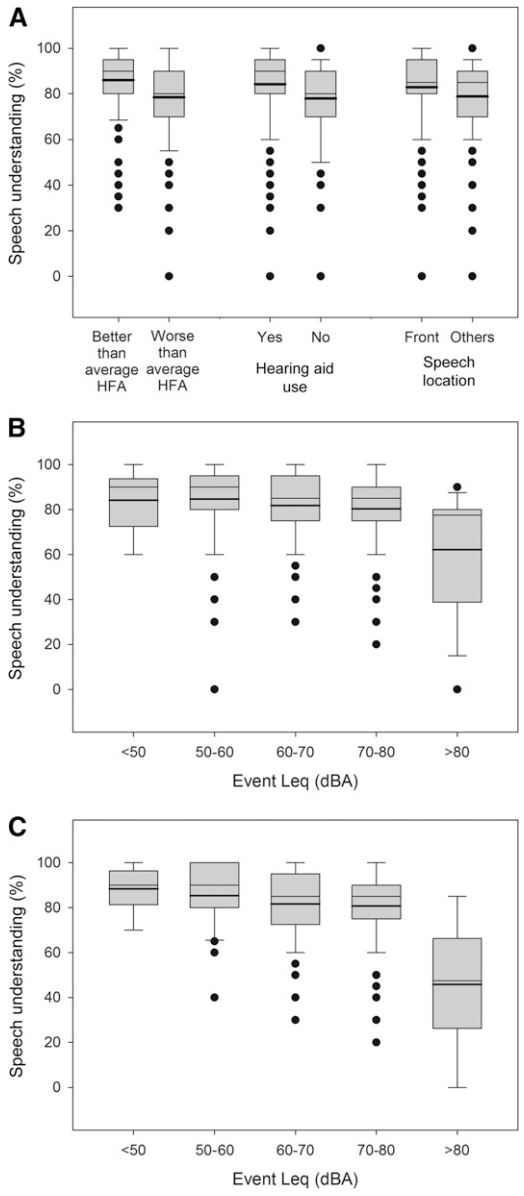

Figure 6. Box plot of self-reported speech understanding as a function of hearing threshold, hearing aid use, primary talker location (A) and event Leq measured using dosimeters (B and C). Panel B includes all surveys involving conversations or speech listening, while panel C excludes surveys that involved phone conversation and media listening. The thinner and thicker lines within the box mark the median and mean, respectively. Error bars indicate the 10th and 90th percentiles. Solid circles are outlying data points.

This is probably due to the weaker association between the dosimeter's overall sound level and the SNR during phone conversations (the overall sound level did not include the speaker's voice on the phone) and media/TV speech listening (when both speech and the noise sounds were from the TV, higher TV volume did not indicate better SNR). If the surveys that involved phone conversation and media listening were eliminated, the trend of speech understanding increasing as event Leq decreased became much clearer (Figure 6C). For the data that did not contain phone conversations and media listening events (n = 386), the mixed model revealed the same results: HFA, hearing aid use, speech location, and event Leq all had a significant effect on speech understanding (Model B in Table 1).

Noisiness Rating

Figure 7 shows a box plot of event Leq as a function of noisiness rating (n = 651). Although the variation was large, noisiness rating increased monotonically as event Leq increased. As in Experiment 1, ordinal logistic regression analysis was performed to determine the relationship between noisiness rating and event Leq. The effect of HFA and hearing aid use was controlled for in the analysis because these two factors might affect the perception of noisiness. This result indicated that the effect of event Leq was significant ($\chi_1^2 = 15.72$, $p < 0.001$), suggesting that the participants tended to report higher noisiness ratings in environments wherein the overall sound levels were higher. The effects of HFA and hearing aid use were not significant.

Lifestyle

To quantify participant's lifestyle using the EMA data, the number of different types of events recorded in all surveys (including the ones that did not involve speech listening) was counted for each participant. It was assumed that those who reported more different types of events (higher event counts) would have more active or diverse lifestyles. To aggregate the dosimeter data for an individual, the event Leq was averaged across all events of a given participant weighted by the event duration. To quantify the variability of environmental sound level, two variables were derived. Between-event variability was the standard deviation

Table 1. Mixed Models on Predicting Self-reported Speech Understanding

Model	Variable	*F* Value	*p* Value
A (n = 667)	HFA	11.6	<0.001
	Hearing aid use	21.59	<0.001
	Primary talker location	8.71	0.003
	Event Leq	23.0	<0.001
B (n = 386)	HFA	7.51	0.006
	Hearing aid use	17.62	<0.001
	Primary talker location	4.66	0.03
	Event Leq	40.5	<0.001

Note: Model A includes all surveys involving conversations or speech listening. Model B excludes surveys that involved phone conversation and media listening.

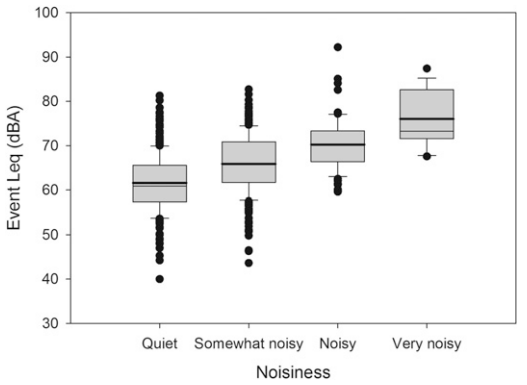

Figure 7. Box plot of event Leq measured using dosimeters as a function of self-reported noisiness. The thinner and thicker lines within the box mark the median and mean, respectively. Error bars indicate the 10th and 90th percentiles. Solid circles are outlying data points.

of the event Leq across all entries of a given participant. To derive the within-event variability, the standard deviation of 5-sec Leqs of a given event was first computed. The standard deviations of all entries of a given participant were then averaged.

The relationships between time-weighted mean Leq, between- and within-event variability, and lifestyle quantified by event count were examined using Pearson's correlations. Although the event count did not correlate to between- ($r = 0.24, p = 0.22$) and within-event variability ($r = -0.003, p = 0.99$), event count was positively associated with mean Leq ($r = 0.45$, $p = 0.018$). Figure 8 shows time-weighted mean Leq as a function of event count. If the outlier indicated by the arrow in the figure was eliminated, the correlation coefficient increased to 0.54 ($p = 0.004$).

Discussion

Event count, which presumably quantified the participant's lifestyle, did not correlate to either between- or within-event variability. This finding is in conflict with the study by Gatehouse et al (2006) that demonstrated the association between auditory lifestyle assessed using a retrospective self-report and the variation of sound level recorded by dosimeters. The reason for this discrepancy is unclear. One possible explanation is that the event count used in this experiment and the retrospective self-report used in Gatehouse et al (2006) assess different aspects of lifestyle.

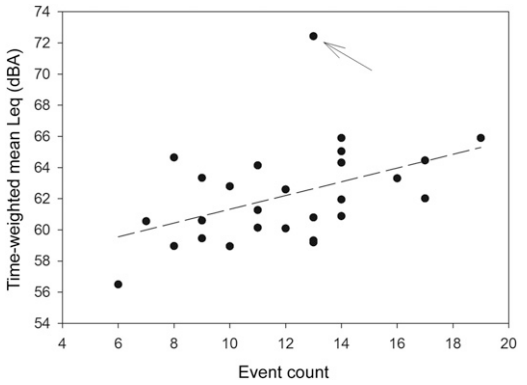

Figure 8. Time-weighted mean Leq as a function of event count. Dashed line represents the regression line. Arrow indicates an outlier.

Regardless, most relationships examined in Experiment 2 were consistent with the established knowledge or theory in audiology: better speech understanding was associated with lower (better) HFA, the use of hearing aids, front-located speech, and lower event Leq; higher noisiness rating was associated with higher event Leq; higher event count (more diverse lifestyle) was associated with higher time-weighted mean Leq. These results suggested that, aggregating data from multiple assessments conducted in a wide range of uncontrolled real-world environments, EMA could generate valid results regarding human listening experiences and relationships between experience and listening context.

GENERAL DISCUSSION

The two experiments presented in this article were designed to examine the construct validity of the EMA methodology in audiology. At the micro level, Experiment 1 suggested that the participants were able to estimate their listening experiences (i.e., speech understanding) and characterize listening context in complicated laboratory settings and in semicontrolled real-world conversations with reasonable accuracy. At the macro level, Experiment 2 indicated that the pattern of the data aggregated across multiple assessments conducted in a wide range of uncontrolled real-world environment was consistent with the established knowledge regarding the relationships between speech understanding, hearing loss, hearing aid use, listening context (talker location and noisiness), and lifestyle. Taken together, the two experiments suggested that, in terms of speech understanding and related listening contexts, EMA reflects what it is intended to measure, supporting the construct validity of EMA in audiology research.

Although this study supported the construct validity of EMA, more research is needed in the future to further validate and optimize this methodology. For example, test–retest reliability is a necessary, although not sufficient, requirement for establishing the validity of a measure. Literature in psychology and sociology has indicated that, when EMA data are aggregated, individuals show a pattern of responses that is consistent with future or past patterns (for a review, see Hektner et al, 2007). The test–retest reliability of EMA in audiology research, however, has not been investigated. Another example for future research is related to the questions

used in EMA. In most previous audiology research, EMA questions were created specifically for the study and, therefore, their wordings and response formats were not vigorously validated. It would be beneficial to establish and validate a set of standardized questions that can be used in EMA. Finally, it has been suggested that EMA data could be useful for clinicians to understand patients' specific communication needs, optimize hearing aid fitting, and provide individualized aural rehabilitation training (Galvez et al, 2012). However, the current format of EMA is not suitable for clinical use due to its high levels of respondent load (Kahneman et al, 2004). Furthermore, systems or models that can convert raw EMA data to meaningful information for clinicians do not exist presently. More research is needed to optimize EMA for clinical use and to empirically determine the value of EMA in clinical settings.

CONCLUSION

EMA has the potential to become an important measure in audiology research. Because EMA can record detailed information about experiences and related contexts from moment to moment, EMA is suited to characterize individuals' listening experiences that are highly affected by physical (e.g., noise level) or social contexts (e.g., talker familiarity). Because the effects of many modern hearing aid features (e.g., directional microphones) are context-dependent, EMA is also suited to assess hearing aid outcomes. The two experiments of this study suggested that (a) adults with hearing impairment were able to report their listening experiences and related contexts and (b) the pattern of their reports collected from a wide range of real-world environments was consistent with the established knowledge, supporting the construct validity of EMA.

REFERENCES

Banerjee S. (2011) Hearing aids in the real world: typical automatic behavior of expansion, directionality, and noise management. *J Am Acad Audiol* 22(1):34–48.

Bentler R, Wu YH, Kettel J, Hurtig R. (2008) Digital noise reduction: outcomes from laboratory and field studies. *Int J Audiol* 47(8):447–460.

Bradburn NM, Rips LJ, Shevell SK. (1987) Answering autobiographical questions: the impact of memory and inference on surveys. *Science* 236(4798):157–161.

Cienkowski KM, Speaks C. (2000) Subjective vs. objective intelligibility of sentences in listeners with hearing loss. *J Speech Lang Hear Res* 43(5):1205–1210.

Cord MT, Walden BE, Surr RK, Dittberner AB. (2007) Field evaluation of an asymmetric directional microphone fitting. *J Am Acad Audiol* 18(3):245–256.

Cox RM. (2003) Assessment of subjective outcome of hearing aid fitting: getting the client's point of view. *Int J Audiol* 42(1, Suppl): S90–S96.

Cox RM, Alexander GC. (1995) The abbreviated profile of hearing aid benefit. *Ear Hear* 16(2):176–186.

Cox RM, Alexander GC, Gilmore C. (1987) Development of the Connected Speech Test (CST). *Ear Hear* 8(5, Suppl): 119S–126S.

Cox RM, Alexander GC, Rivera IM. (1991) Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners. *J Speech Hear Res* 34(4):904–915.

Cronbach LJ, Meehl PE. (1955) Construct validity in psychological tests. *Psychol Bull* 52(4):281–302.

Galvez G, Turbin MB, Thielman EJ, Istvan JA, Andrews JA, Henry JA. (2012) Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users. *Ear Hear* 33(4):497–507.

Gatehouse S, Naylor G, Elberling C. (2006) Linear and nonlinear hearing aid fittings—2. Patterns of candidature. *Int J Audiol* 45(3): 153–171.

Hasan SS, Chipara O, Wu YH, Aksan N. (2014) Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones. *Pervasive Computing Technologies for Healthcare*, 126–133.

Hasan SS, Lai F, Chipara O, Wu YH. (2013) AudioSense: enabling real-time evaluation of hearing aid technology in-situ. *International Symposium on Computer-Based Medical Systems (CBMS)*, 143–148.

Hektner JM, Schmidt JA, Csikszentmihalyi M. (2007) *Experience Sampling Method: Measuring the Quality of Everyday Life*. Thousand Oaks, CA: Sage.

Henry JA, Galvez G, Turbin MB, Thielman EJ, McMillan GP, Istvan JA. (2012) Pilot study to evaluate ecological momentary assessment of tinnitus. *Ear Hear* 33(2):179–290.

Humes LE. (2002) Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *J Acoust Soc Am* 112(3 pt 1):1112–1132.

Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA. (2004) A survey method for characterizing daily life experience: the day reconstruction method. *Science* 306(5702):1776–1780.

Lee L, Lau C, Sullivan D. (1998) The advantage of a low compression threshold in directional microphones. *Hear Rev* 5:30–32.

Lunner T. (2003) Cognitive function in relation to hearing aid use. *Int J Audiol* 42(1, Suppl):S49–S58.

Palmer C, Bentler R, Mueller HG. (2006) Evaluation of a second-order directional microphone hearing aid: II. Self-report outcomes. *J Am Acad Audiol* 17(3):190–201.

Pearsons KS, Bennett RL, Fidell S. (1976) *Speech Levels in Various Environments. Report to the Office of Resources and Development*. Cambridge, MA: Bolt, Beranek and Newman.

Preminger JE, Cunningham DR. (2003) Case-study analysis of various field study measures. *J Am Acad Audiol* 14(1): 39–55.

Ricketts T, Henry P, Gnewikow D. (2003) Full time directional versus user selectable microphone modes in hearing aids. *Ear Hear* 24 (5):424–439.

Ricketts TA, Hornsby BW. (2003) Distance and reverberation effects on directional benefit. *Ear Hear* 24(6):472–484.

Shiffman S, Stone AA, Hufford MR. (2008) Ecological momentary assessment. *Annu Rev Clin Psychol* 4:1–32.

Studebaker GA. (1985) A "rationalized" arcsine transform. *J Speech Hear Res* 28(3):455–462.

Sumby WH, Pollack I. (1954) Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 26:212–215.

Walden BE, Surr RK, Cord MT, Dyrlund O. (2004) Predicting hearing aid microphone preference in everyday listening. *J Am Acad Audiol* 15(5):365–396.

Walden BE, Surr RK, Grant KW, Van Summers W, Cord MT, Dyrlund O. (2005) Effect of signal-to-noise ratio on directional microphone benefit and preference. *J Am Acad Audiol* 16(9):662–676.

Wu YH, Bentler RA. (2010a) Impact of visual cues on directional benefit and preference: part I—laboratory tests. *Ear Hear* 31(1):22–34.

Wu YH, Bentler RA. (2010b) Impact of visual cues on directional benefit and preference: part II—field tests. *Ear Hear* 31(1):35–46.

Wu YH, Bentler RA. (2012) Do older adults have social lifestyles that place fewer demands on hearing? *J Am Acad Audiol* 23(9): 697–711.

Wu YH, Stangl E. (2013) The effect of hearing aid signal-processing schemes on acceptable noise levels: perception and prediction. *Ear Hear* 34(3):333–341.

Wu YH, Stangl E, Bentler RA. (2013) Hearing-aid users' voices: a factor that could affect directional benefit. *Int J Audiol* 52(11):789–794.