Test-Retest Reliability of Ecological Momentary Assessment in Audiology Research

Yu-Hsiang Wu, MD, PhD[1] Elizabeth Stangl, AuD[1] Octav Chipara, PhD[2] Xuyang Zhang, PhD[1]

[1] Department of Communication Sciences and Disorders, The University of Iowa, Iowa City, Iowa
[2] Department of Computer Science, The University of Iowa, Iowa City, Iowa

Address for correspondence Yu-Hsiang Wu, MD, PhD, Department of Communication Sciences and Disorders, The University of Iowa, 125C WJSHC, Iowa City, IA 52242 (e-mail: yu-hsiang-wu@uiowa.edu).

J Am Acad Audiol

Abstract

Background Ecological momentary assessment (EMA) is a methodology involving repeated surveys to collect in situ data that describe respondents' current or recent experiences and related contexts in their natural environments. Audiology literature investigating the test-retest reliability of EMA is scarce.

Purpose This article examines the test-retest reliability of EMA in measuring the characteristics of listening contexts and listening experiences.

Research Design An observational study.

Study Sample Fifty-one older adults with hearing loss.

Data Collection and Analysis The study was part of a larger study that examined the effect of hearing aid technologies. The larger study had four trial conditions and outcome was measured using a smartphone-based EMA system. After completing the four trial conditions, participants repeated one of the conditions to examine the EMA test-retest reliability. The EMA surveys contained questions that assessed listening context characteristics including talker familiarity, talker location, and noise location, as well as listening experiences including speech understanding, listening effort, loudness satisfaction, and hearing aid satisfaction. The data from multiple EMA surveys collected by each participant were aggregated in each of the test and retest conditions. Test-retest correlation on the aggregated data was then calculated for each EMA survey question to determine the reliability of EMA.

Results At the group level, listening context characteristics and listening experience did not change between the test and retest conditions. The test-retest correlation varied across the EMA questions, with the highest being the questions that assessed talker location (median $r = 1.0$), reverberation ($r = 0.89$), and speech understanding ($r = 0.85$), and the lowest being the items that quantified noise location (median $r = 0.63$), talker familiarity ($r = 0.46$), listening effort ($r = 0.61$), loudness satisfaction ($r = 0.60$), and hearing aid satisfaction ($r = 0.61$).

Conclusion Several EMA questions yielded appropriate test-retest reliability results. The lower test-retest correlations for some EMA survey questions were likely due to fewer surveys completed by participants and poorly designed questions. Therefore, the present study stresses the importance of using validated questions in EMA. With sufficient numbers of surveys completed by respondents and with appropriately designed survey questions, EMA could have reasonable test-retest reliability in audiology research.

Keywords
- ecological momentary assessment
- hearing aid
- outcome

received
September 23, 2019
accepted after revision
February 10, 2020

DOI https://doi.org/10.1055/s-0040-1717066.
ISSN 1050-0545.

Ecological momentary assessment (EMA), also known as experience sampling or ambulatory assessment, is a methodology that asks respondents to repeatedly report their experiences during or shortly after the experiences in their natural environments (i.e., in situ self-reports).[1] Because EMA can provide a rich description of a sample of moments in respondents' lives while avoiding the distortions that affect the delayed recall and evaluation of experiences, EMA is considered to be less affected by recall bias compared with retrospective questionnaires. Because detailed contextual information can also be collected in each assessment, EMA could have high contextual resolution. Therefore, the use of EMA in audiology research has grown in the past decades. EMA has been implemented using paper-and-pencil journals,[2–5] daily diaries,[6] portable computers,[7] and smartphones[8,9] to assess hearing difficulty or hearing aid outcomes in the real world.

Although the use of EMA in audiology research is increasing, audiology literature examining the psychometric characteristics of this methodology, such as construct validity and test-retest reliability, is scarce. Wu et al[10] conducted a field study to determine the construct validity of EMA in audiology research. These researchers found that the pattern of the self-reported data aggregated across multiple EMA surveys conducted in a wide range of uncontrolled real-world environment was consistent with the established knowledge in audiology. For example, better speech understanding reported in EMA surveys was associated with lower (better) hearing thresholds, the use of hearing aids, and front-located speech. Furthermore, higher noisiness level (i.e., noisier) reported in EMA surveys was associated with higher sound level measured using noise dosimeters. Based on these results, Wu et al suggested that, regarding speech understanding and related listening contexts, EMA reflects what it is intended to measure, supporting its construct validity. Timmer et al[11] examined the feasibility and construct validity of EMA. Because of the close agreement between self-reported listening context characteristics and objective data from the hearing aid classifier, Timmer et al[11] concluded that EMA is a valid research methodology.

Another important consideration for any instrument that might be used to determine intervention outcomes is test-retest reliability. In the literature, test-retest reliability of an instrument is often quantified by the correlation between test and retest conditions. For retrospective questionnaires that are widely used in audiology research, their test-retest reliabilities vary considerably: Hearing Handicap Inventory for the Elderly (HHIE): 0.79 to 0.98 (paper-and-pencil administration)[12]; Abbreviated Profile of Hearing Aid Benefit (APHAB): 0.65 to 0.89[13]; Satisfaction with Amplification in Daily Life (SADL): 0.52 to 0.81[14]; International Outcome Inventory for Hearing Aids: 0.62 to 0.73[15]; Speech, Spatial, and Qualities hearing scale (SSQ): 0.56 to 0.83 (paper-and-pencil administration).[16] It has been suggested that a test-retest correlation higher than 0.7 to 0.8 is desirable.[17,18]

Examining the test-retest correlation of EMA, however, is less straightforward. Contrary to one-time measures such as retrospective questionnaires, the EMA does not assume that people will have entirely consistent responses in every EMA survey. That is, the reliability of EMA responses is expected to vary as a person's experiences and contexts do not remain stagnant over time. Because perfect test-retest reliability across EMA surveys is unlikely to occur, previous research typically used aggregated EMA survey data to determine the test-retest reliability. It is assumed that EMA data aggregated from multiple surveys completed by an individual will show a uniform response pattern across time.[19] As reviewed by Csikszentmihalyi and Larson,[20] the test-retest reliability of EMA varies considerably, ranging from 0.38 to 0.77. Note that some of the lower test-retest correlations were due to previous studies often splitting 1-week data to examine the association between average ratings from the first and second halves of the week. The reliability of EMA data could be susceptible to systematic variations in weekly schedules.[19]

Few researchers have examined the test-retest reliability of EMA in audiology research. In a case study by Preminger and Cunningham,[2] eight participants were asked to use paper-and-pencil journals to rate the speech clarity of hearing aids. Two hearing aid gain frequency responses were used, with one supposed to generate better speech clarity than the other. For each gain frequency response, study participants rated the speech clarity on a 10-point scale (i.e., the scale rating technique) three times per day in the journals for 1 week (i.e., the test condition). The whole procedure was repeated weeks later (the retest condition). In addition, the paper-and-pencil EMA was used with a paired comparison technique in another set of test-retest conditions, in which the two gain frequency responses were saved into two programs of the hearing aids. Participants switched between the two programs and then recorded in the journals which one provided better speech clarity. Statistical analysis was conducted for each participant. For the EMA condition with the scale rating technique, only four out of the eight participants were able to report the correct results (i.e., reporting higher clarity ratings for the gain frequency response that was supposed to generate better speech clarity) and only one of the four participants had consistent results across the test and retest conditions. For the EMA condition with the paired comparison technique, six out of the eight participants reported the correct results and five of these six participants had consistent results between the test and retest conditions. Therefore, Preminger and Cunningham[2] concluded that the EMA methodology with the scale rating technique was less reliable than the paired comparison technique. Because Preminger and Cunningham[2] conducted the statistical analysis at the individual level, no test-retest correlation was reported.

The objective of the present study was to determine the test-retest reliability of EMA in measuring listening context and listening experience for older hearing aid users. Both the test and retest condition lasted seven consecutive days to minimize the systematic variations in weekly schedules. The data from multiple EMA surveys collected by each participant were aggregated in each of the test and retest conditions. Test-retest correlation on the aggregated data was

then calculated to determine the reliability of EMA. It was expected that the self-reported data would vary considerably across individual EMA surveys. It was also expected that, when multiple EMA survey data were aggregated, the variation of the data would be minimized. Therefore, it was hypothesized that the test-retest reliability of EMA would be similar to that of retrospective questionnaires.

Methods

The present study was part of a larger study designed to examine the effect of advanced hearing aid noise reduction technologies.[9] Older adults with hearing loss were fitted with bilateral hearing aids. Two hearing aid models, one a less expensive, basic-level device (basic hearing aid) and the other a more expensive, advanced-level device (premium hearing aid), were used. The noise reduction technologies (directional microphones and noise reduction algorithms) of the hearing aids were turned on or off to create four trial conditions. A single-blinded, crossover repeated measures design was used. The participants were blinded to the technology level and settings of the hearing aids and all aids were physically identical. During the field trial of each condition, the participants wore the hearing aids in their daily lives for 5 weeks. After completing all four trial conditions, each participant repeated one of the four conditions (randomly selected, see below for details) to examine the test-retest reliability of the EMA methodology.

Participants

Fifty-one participants (25 males and 26 females) recruited from cities, towns, and farms around eastern Iowa and north western Illinois completed the study. Their ages ranged from 65 to 88 years with a mean of 73.7 years. Participants were eligible for inclusion in the study if their hearing loss met the following criteria: (1) postlingual, bilateral, sensorineural hearing loss (air-bone gap < 10 dB); (2) pure-tone average across 0.5, 1, 2, and 4 kHz greater than 25 dB hearing level (HL) but not worse than 60 dB HL; and (3) hearing symmetry within 20 dB for all test frequencies. The mean pure-tone thresholds are shown in ►**Fig. 1**. All participants were native English speakers. Upon entering the study, 30 participants had at least 1 year of previous hearing aid experience. No previous smartphone experience was necessary for inclusion in the study.

Hearing Aids and Fitting

Two commercially available behind-the-ear hearing aid models were used in the larger study: basic (retail price per pair $\approx$ \$1,500) and premium (retail price per pair $\approx$ \$5,000) hearing aids. Devices were coupled to the participant's ears bilaterally using slim tubes and custom canal earmolds with clinically appropriate vent sizes. The devices were programmed to meet real-ear aided response targets ($\pm$3 dB) specified by the second version of the National Acoustic Laboratory nonlinear prescriptive formula (NAL-NL2[21]) and was verified using a probe-microphone hearing aid analyzer (Audioscan Verifit; Dorchester, Ontario, Canada)

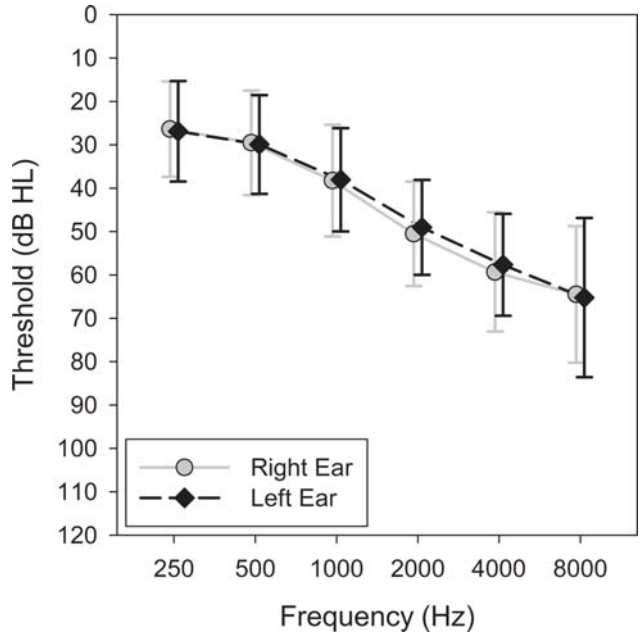

Fig. 1 Average audiograms for left and right ears of study participants. Error bars = 1 standard deviation (SD).

with a 65-dB sound pressure level speech signal presented from 0-degree azimuth. The status of noise reduction technologies of basic and premium hearing aids was manipulated to create four trial conditions: basic-on, basic-off, premium-on, and premium-off (2×2 factorial design). See Wu et al[9] for the details of the technologies. All other features (e.g., wide dynamic range compression, adaptive feedback suppression, and low-level expansion) remained active at default settings. The volume control of the device was disabled during the trial.

EMA

The EMA methodology was used to collect the participant's in situ self-reports. The EMA was implemented using Samsung (Seoul, South Korea) Galaxy S3 smartphones (i.e., smartphone-based EMA). Smartphone application software (i.e., app) was developed to deliver electronic surveys.[22] The app prompted participants to complete surveys at randomized intervals approximately every 2 hours within a participant's specified time window. The 2-hour interprompt interval was selected because it seemed to be a reasonable balance between participant burden, compliance, and the amount of data that would be collected.[23] Participants were instructed to answer survey questions based on their experiences during the past 5 minutes so that recall bias was minimized. Participants could also initiate a survey to report a listening event. For the latter type of surveys, although participants were encouraged to complete the surveys during the listening event, they were allowed to report the experience up to 1 hour after the event.

In each EMA survey, a series of questions regarding listening context and listening experience were asked. See ►**Table 1** for the survey questions and the associated responses. For example, the survey first asked "*Were you listening to speech?*" and provided two options for the

Table 1 EMA survey questions and responses

Questions	Responses
1. Were you listening to speech?	☐ Yes ☐ No
1a. [Activity] (If "Yes") What were you listening to?	☐ Conversation, 3 or fewer ☐ Conversation, 4 or more ☐ Speech listening, live ☐ Speech listening, media ☐ Conversation, phone
1b. [Activity] (If "No") What were you listening to?	☐ Non-speech sound listening ☐ Not actively listening
2. Where were you?	☐ Outdoor/Traffic ☐ Indoor
2a. [Location] (If "Outdoor/Traffic") Please be more specific.	☐ Outdoor, moving traffic ☐ Outdoor, other than traffic
2b. [Location] (If "Indoor") Please be more specific.	☐ Home, 10 or fewer ☐ Other than home, 10 or fewer ☐ Crowd of people, 11 or more
3. [Talker Familiarity] (If listening to speech) Were you familiar with the talker(s)?	☐ Unfamiliar ☐ Somewhat unfamiliar ☐ Somewhat familiar ☐ Familiar
4. [Visual Cues] (If listening to speech) Could you see the talker's face?	☐ No ☐ Yes, but only sometime ☐ Almost always
5. [Talker Location] (If listening to speech, but not on the phone) Where was the talker?	☐ Front ☐ Side ☐ Back
6. [Noisiness] On average, how noisy was it?	☐ Quiet ☐ Somewhat noisy ☐ Noisy ☐ Very noisy
7. [Noise Location] (If not quiet) Where was the noise?	☐ Front ☐ Side ☐ Back ☐ All around
8. (If indoor) Compared with an average living room, how large was the room?	☐ Smaller ☐ About average ☐ Larger
9. (If indoor) Was there carpeting?	☐ Yes ☐ No
10. [Speech Understanding] (If listening to speech) On average, how much speech did you understand?	0% — 100%
11. [Listening Effort] (If active listening) On average, how much effort was required to listen effectively?	Very easy — Very effortful
12. [Loudness] How would you judge the level of loudness of the sound?	Very soft — Comfortable — Uncomfortably loud
13. [Loudness Satisfaction] Were you satisfied with the loudness?	Not good at all — Just right
14. [Localization] In general, could you tell where sounds were coming from right away?	Not at all — Perfectly
15. Did you use your hearing aids?	☐ Yes ☐ No
16. [Hearing Aid Satisfaction] (If using hearing aids) Were you satisfied with your hearing aids?	Not at all — Very satisfied

Abbreviation: EMA, ecological momentary assessment.

participants to select (yes/no). Participants tapped a button on the smartphone screen to indicate their responses. Next, the survey asked a question about the listening activity ("*What were you listening to?*") and provided five response options (if the answer to the previous question was yes) or two options (if the answer to the previous question was no). In the data analysis of the present study, this variable was referred to as the Activity variable (square brackets in ►**Table 1**) and had seven possible responses. The survey then presented more questions to assess the characteristics of the listening context, including location, talker familiarity, availability of visual cues, talker location, noisiness, noise location, room size, and carpeting. Note that the questions about room size and carpeting, plus the question that asked outdoor versus indoor (Question 2 in ►**Table 1**), were used to estimate the reverberation (i.e., the Reverberation variable). According to Walden et al,[3] outdoors and traffic were assumed to be low-reverberant environments. Indoor, carpeted spaces that were equal in size or smaller than an average living room were considered to be low-reverberant environments. The remaining indoor locations were assumed to be high-reverberant. Also note that the variables describing the listening context were either categorical variables (Activity, Location, Talker Location, and Noise Location) or ordinal variables (Talker Familiarity, Visual Cues, Noisiness, and Reverberation).

Next, the app presented a series of questions to assess the participant's listening experiences, including speech understanding, listening effort, perceived loudness, loudness satisfaction, sound localization, and hearing aid satisfaction. Because these variables served as the outcomes of the larger study, visual analog scales (rather than the categorical or ordinal scales) were used to obtain fine-grained information. See ►**Table 1** for the anchors used in the scales. Participants used the sliding bar on the visual analog scale to mark their perception. The participant's rating was quantified by multiplying the ratio of the distance between the left end of the scale and the participant's mark to the length of the entire scale by 100. All variables that described listening experience were interval variables.

Note that although Noisiness (Question 6 in ►**Table 1**) and Loudness (Question 12) might assess similar constructs, the former was classified as a listening context variable while the latter was treated as a listening experience variable in the present study. The reason for this was somewhat arbitrary. Noisiness was a context variable because it has been shown to be associated with signal-to-noise ratio.[24] On the other hand, because Loudness and Loudness Satisfaction (Questions 12 and 13) were adapted from the Profile of Aided Loudness questionnaire,[25] they were viewed as listening experience variables. The participants were instructed to answer these two questions based on the loudness of the overall sound. Also note that the EMA survey questions shown in ►**Table 1** were created specifically for the larger study and, therefore, their wordings and response formats have not been vigorously validated.

The survey questions were presented adaptively such that certain answers determined whether follow-up questions would be elicited. The presentation logic is shown in parentheses in ►**Table 1**. For example, the noise location question would not be presented if the participant indicated "Quiet" in the noisiness question. The speech understanding question would be presented only when participants indicated that they were listening to speech in the beginning of the survey.

Retrospective Questionnaire

To compare the test-retest reliability of EMA and retrospective questionnaire, four widely used instruments were administered.

Hearing Handicap Inventory for the Elderly

The HHIE[26] is a 25-item inventory designed to evaluate the emotional and social impact of hearing loss on an individual's life. The HHIE is divided into two subscales: the Emotional subscale, which measures how emotional responses in an individual's life are influenced by hearing loss, and the Social subscale, which assesses the extent to which social aspects of an individual's life are impacted by hearing loss.

Abbreviated Profile of Hearing Aid Benefit

The APHAB[13] is a 24-item inventory designed to evaluate benefit experienced from hearing aid use and to quantify the degree of communication difficulty experienced in various situations. The questionnaire consists of four subscales. The Ease of Communication, Reverberation, and Background Noise subscales are focused on speech communication. The Aversiveness subscale evaluates the individual's response to unpleasant environmental sounds.

Satisfaction with Amplification in Daily Life

The SADL[14] is a 15-item inventory designed to evaluate an individual's satisfaction with his/her hearing aids. The questionnaire is divided into four subscales. The Positive Effect subscale quantifies improved performance while using hearing aids. The Service and Cost subscale measures the adequacy of service provided by the professional and the cost of the devices. The Negative Features subscale assesses undesirable aspects of hearing aid use. The Personal Image subscale evaluates the domain of self-image and stigma. Note that in the present study the Service and Cost subscale contained only one item. This is because the other two items in this subscale are related to hearing aid cost and repair and therefore were not applicable to the study (hearing aids were provided at no cost in the study).

Speech, Spatial, and Qualities Hearing Scale

The 49-item SSQ[27] is a questionnaire designed to measure a range of hearing disabilities across several domains. The SSQ consists of three subscales that measure the ability of an individual to understand speech (Speech subscale), to localize acoustic events (Spatial subscale), and to evaluate auditory experience including music perception and the clarity and naturalness of sound (Qualities subscale).

Procedures

The study was approved by the Institutional Review Board at The University of Iowa. After signing the consent form,

participants' hearing thresholds were measured using pure-tone audiometry. If participants met all the inclusion criteria, earmold impressions were taken by an audiologist. Next, demonstrations of how to work and care for the smartphone, as well as taking and initiating EMA surveys on the phone, were provided. Participants were instructed to respond to the auditory/vibrotactile prompts to take surveys whenever it was possible and within reason (e.g., not while driving). Participants were also encouraged to initiate a survey during or immediately after they had encountered a different listening experience lasting at least 10 minutes. Each participant was given a set of take-home written instructions detailing how to use and care for the phone, as well as when and how to take the surveys. Once all the participants' questions had been answered and they demonstrated competence in the ability to perform all the related tasks, they were sent home with one smartphone and began a 3-day practice session. Participants returned to the laboratory after the practice session. If participants misunderstood any of the EMA/smartphone-related tasks during the practice session, they were reinstructed on how to properly use the equipment or take the surveys.

Next, the hearing aids of all four trial conditions (basic-on, basic-off, premium-on, and premium-off) were fit and the first field trial condition began. The order of the four trial conditions was randomized across participants. In each trial condition, participants familiarized themselves with the hearing instrument settings for 4 weeks. Participants then returned to the laboratory and the outcome measures of the larger study were conducted. Participants were then given smartphones and the assessment week in which participants carried smartphones to conduct EMA surveys began. Participants were encouraged to go about their normal daily routines during the week. One week later, participants brought the smartphones back to the laboratory and the data saved in the smartphone were downloaded. Retrospective questionnaires were administered. Hearing aids were inspected, cleaned, and reprogrammed. Then, the next trial condition began. See Wu et al[9] for more details on the larger study. After participants completed all four trial conditions, participants repeated one of the trial conditions to examine the test-retest reliability of EMA (4 weeks wearing hearing aids plus 1 week for EMA). Specifically, for each participant, the hearing aid model that was used in the last (i.e., the fourth) trial condition was used in the retest condition. However, the status of noise reduction features (on vs. off) was randomly selected for each participant. This design (rather than using the same hearing aid model and configuration across all participants in the retest condition) was used because of the limited number of hearing aids available to the larger study. In the present article, the two conditions in which the same hearing aid model and configuration were used were referred to as the test and retest conditions, respectively. It was determined a priori that the data from all hearing aid models and configurations would be pooled together for analysis. Pooling the data obtained under rather different hearing aid conditions would make the findings of the present study more generalizable than had they been obtained under just a single condition. Monetary compensation was provided to the participants upon completion of the study.

Results

The mean interval between the test and retest condition was 8.7 weeks (standard deviation [SD] = 4.2), ranging from 4 to 27 weeks. Across the test and retest conditions, 3,900 EMA surveys were completed by the 51 participants (test: $n = 1{,}995$; retest: $n = 1{,}905$). On average, each participant completed 39.1 surveys (SD = 14.4, range = 9–74, median = 37) and 37.4 surveys (SD = 15.0, range = 4–80, median = 37) in the test and retest conditions, respectively. Among all surveys, 2,648 (68%) were prompted by the EMA app (test: $n = 1{,}331$; retest: $n = 1{,}317$) and the remaining 1,252 (32%) were initiated by the participants (test: $n = 664$; retest: $n = 588$). Although it would be interesting to examine the reliability of app-initiated and participant-initiated surveys separately, it was determined a priori that both types of surveys would be pooled together to ensure that there were sufficient numbers of surveys from each participant for analysis. Further, it is not uncommon for EMA studies to pool data from both types of surveys to answer research questions.[8,9,24]

Comparisons: Test versus Retest

Data were examined to determine if the listening context and listening experience was stable across the test and retest conditions at the group level. To this end, EMA data were first aggregated within each participant. For the categorical and ordinal variables that described listening contexts, the probability of each response being selected in each of the test and retest conditions was calculated for each participant. For the interval variables that described listening experiences, the ratings obtained from the visual analog scale were averaged across individual EMA surveys for each participant in each condition. ►**Tables 2** and **3** show the mean probability (categorical and ordinal variables) and rating (interval variables) averaged across the participants, respectively, of the test and retest conditions. The mean probability and mean ratings were quite similar across the test and retest conditions.

Statistical analysis was then conducted on the aggregated data. For the categorical and ordinal variables, repeated measures analysis of variance (ANOVA) was conducted to determine the effect of test-retest condition on the distribution of response probability. The independent variables were response, test-retest condition, and their interaction. The dependent variable was response probability. Separate analysis was conducted for each variable. Because for a given variable the probability for a response is equal to one minus the sum of the probabilities of the remaining responses, including the probability data of all responses in the analysis would violate the ANOVA assumption regarding the independence of samples. Therefore, for all variables except for the Reverberation variable, the last response of a given variable (e.g., "very noisy" of the Noisiness variable) was

Table 2 Mean and standard deviation of the probability of each response being selected of the variable that describes listening context in the test and retest conditions

	Test condition, mean (SD)	Retest condition, mean (SD)	Effect of test-retest condition (*p*-value)	Interaction (*p*-value)
Activity				
Conversation, 3 or fewer	0.26 (0.18)	0.25 (0.18)	0.55	0.19
Conversation, 4 or more	0.07 (0.07)	0.06 (0.06)		
Speech listening, live	0.03 (0.04)	0.02 (0.03)		
Speech listening, media	0.29 (0.22)	0.32 (0.24)		
Conversation, phone	0.05 (0.06)	0.05 (0.07)		
Non-speech sound listening	0.09 (0.15)	0.10 (0.14)		
Not actively listening	0.21 (0.16)	0.22 (0.16)		
Location				
Outdoor, moving traffic	0.07 (0.06)	0.07 (0.06)	0.46	0.02
Outdoor, other than traffic	0.07 (0.10)	0.05 (0.08)		
Indoor, home, 10 or fewer	0.68 (0.16)	0.73 (0.17)		
Indoor, other than home, 10 or fewer	0.11 (0.10)	0.09 (0.09)		
Indoor, crowd of people, 11 or more	0.07 (0.06)	0.06 (0.06)		
Talker Familiarity				
Unfamiliar	0.20 (0.15)	0.20 (0.18)	0.72	0.88
Somewhat unfamiliar	0.12 (0.18)	0.12 (0.14)		
Somewhat familiar	0.22 (0.18)	0.21 (0.20)		
Familiar	0.46 (0.23)	0.48 (0.24)		
Visual Cues				
No	0.23 (0.19)	0.24 (0.18)	0.49	0.31
Yes, but only sometime	0.42 (0.23)	0.39 (0.25)		
Almost always	0.35 (0.23)	0.37 (0.26)		
Talker Location				
Front	0.69 (0.22)	0.69 (0.25)	0.73	0.97
Side	0.27 (0.21)	0.28 (0.22)		
Back	0.04 (0.08)	0.04 (0.05)		
Noisiness				
Quiet	0.58 (0.22)	0.58 (0.22)	0.74	0.94
Somewhat noisy	0.33 (0.20)	0.33 (0.19)		
Noisy	0.07 (0.07)	0.06 (0.09)		
Very noisy	0.02 (0.04)	0.02 (0.04)		
Noise Location				
Front	0.26 (0.28)	0.26 (0.28)	0.55	0.93
Side	0.09 (0.13)	0.10 (0.12)		
Back	0.03 (0.06)	0.03 (0.07)		
All around	0.62 (0.30)	0.61 (0.31)		
Reverberation				
Low	0.53 (0.25)	0.54 (0.27)	0.69	
High	0.47 (0.25)	0.46 (0.27)		

Abbreviations: ANOVA, analysis of variance; SD, standard deviation.
Note: For all variables except for the Reverberation variable, the *p*-values are the results of repeated measures ANOVA that examines the main effect of test-retest condition and the interaction between response and test-retest condition. For the Reverberation variable, the *p*-value is the result of a paired *t*-test that compares the probability of the "Low" response between the test and retest condition. See text for more details.

Table 3 Mean and standard deviation of the rating of the variable that describes listening experience in the test and retest conditions

	Test condition, mean (SD)	Retest condition, mean (SD)	*p*-value
Speech Understanding	82.9 (9.9)	82.8 (9.5)	0.89
Listening Effort	34.4 (16.1)	34.7 (16.2)	0.90
Loudness	44.7 (9.2)	45.1 (9.4)	0.74
Loudness Satisfaction	70.6 (11.2)	69.5 (11.1)	0.42
Localization	79.2 (12.1)	79.3 (12.2)	0.95
Hearing Aid Satisfaction	75.1 (15.1)	71.2 (18.2)	0.07

Abbreviation: SD, standard deviation.
Note: The *p*-values are the results of paired *t*-tests that compare the mean ratings between the test and retest conditions.

not included in the analysis. For these analyses, either the interaction or the main effect of test-retest condition being significant would suggest a difference in the distribution of response probability between the test and retest conditions. Because the Reverberation variable had only two responses ("Low" and "High"), a paired *t*-test was conducted to compare the probability of the "Low" response between the test and retest condition. The results indicate that, except for the interaction of the Location variable ($p = 0.02$; ►**Table 2**), the interactions and the main effect of test-retest condition of all variables were not significant (all $p > 0.05$; ►**Table 2**). These results suggested that the distribution of response probability was similar in the test and retest conditions for most variables that described listening contexts.

For the interval variables that described listening experience, paired *t*-tests were conducted to compare the mean ratings in the test and retest conditions. The results indicated that none of the difference was significant (all $p > 0.05$; ►**Table 3**). Of note is that the mean hearing aid satisfaction rating decreased from 75.1 points in the test condition to 71.2 points in the retest condition, despite that the difference was not statistically significant ($p = 0.07$). The results shown in ►**Tables 2** and **3** suggested that in general the listening context and listening experience was stable over time at the group level.

Test-Retest Reliability

In the literature, test-retest reliability is often examined using test-retest correlation coefficients. Because correlation analysis is more appropriate for interval variables, a different approach was used to determine the test-retest reliability of the categorical variables of the present study (Activity, Location, Talker Location, and Noise Location). In short, the responses to a given question were first ranked from the most to the least frequent responses within each participant and within each test-retest condition. The Spearman's correlation was then calculated between the ranks from the two conditions for each participant. A higher correlation indicated that the order of the frequency of the participant's selection was more consistent between the conditions (e.g., the most common response in the test condition was also the most common response in the retest condition). Next, the correlation coefficients were examined across all participants and boxplots were created. The results are shown in ►**Fig. 2**. The median correlation coefficients of Activity, Location, Talker Location, and Noise Location were 0.75, 0.80, 1.0, and 0.63, respectively. The variation in correlation coefficient across participants was substantial, especially for the Noise Location variable.

For the ordinal variables (Talker Familiarity, Visual Cues, Noisiness, and Reverberation), they were treated as interval variables. Consecutive integers were assigned to the responses based on the order of response (e.g., for the Talker Familiarity variable, Unfamiliar = 1, Somewhat unfamiliar = 2, Somewhat familiar = 3, and Familiar = 4). This approach was similar to the scoring method used by the SADL.[14] The data collected by a given participant were then aggregated by calculating the mean of the response orders for each participant in each condition. Next, Pearson-product moment correlation in the mean response order between the test and retest conditions across all participants was calculated. The scatter plots and the correlation coefficients are shown in ►**Fig. 3**. All correlations were significant ($p < 0.01$), with the highest and lowest correlations being

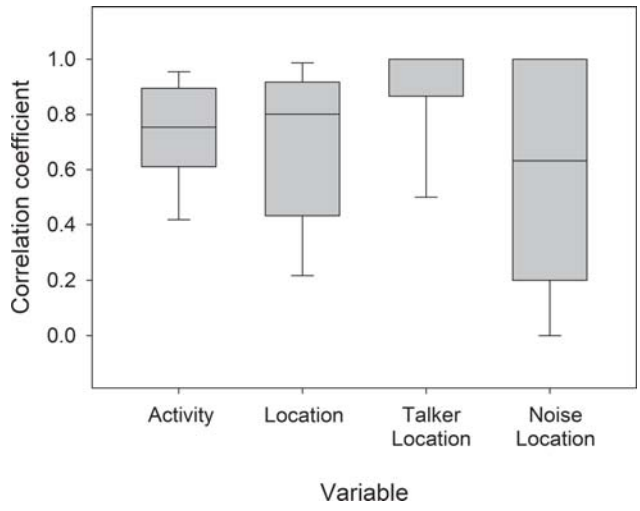

Fig. 2 Boxplots of the test-retest correlation coefficient of the categorical variables that described listening context characteristics. The boundaries of the boxes represent the 25th and 75th percentile and the line within the boxes marks the median. Error bars indicate the 10th and 90th percentiles.

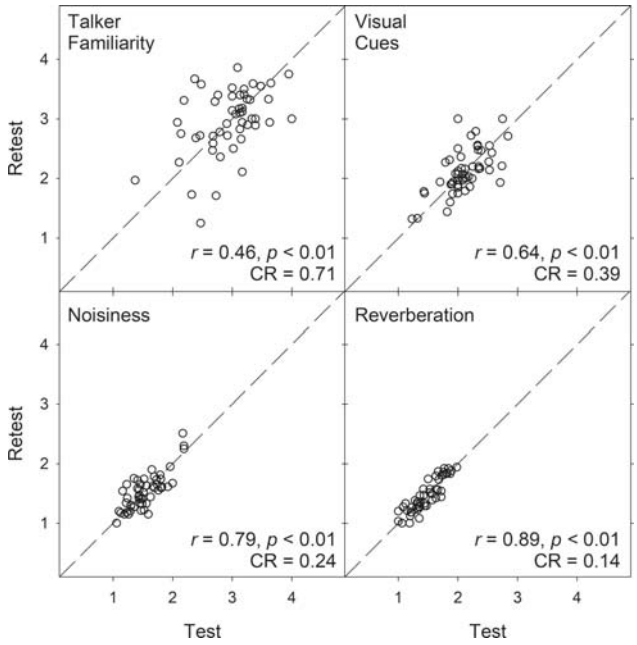

Fig. 3 Scatter plots of the responses of the ordinal variables that described listening context characteristics in the test and retest conditions. Diagonal lines represent perfect match between the test and retest conditions. Talker Familiarity: unfamiliar = 1, somewhat unfamiliar = 2, somewhat familiar = 3, and familiar = 4; Visual Cues: no = 1, yes, but only sometimes = 2, almost always = 3; Noisiness: quiet = 1, somewhat noisy = 2, noisy = 3, very noisy = 4; Reverberation: low = 1, high = 2. CR, coefficient of repeatability.

Reverberation ($r = 0.89$) and Talker Familiarity ($r = 0.46$), respectively. Because it has been suggested that the test-retest correlation coefficient may not be the best way to quantify the repeatability of a measure,[28] the coefficient of repeatability—a value below which the absolute difference between test and retest results would lie with 95% probability—was also calculated (►**Fig. 3**). The coefficient of repeatability shown in ►**Fig. 3**, however, should be interpreted with caution because the score ranges of the four ordinal variables are not identical.

For the interval variables that described listening experience (Speech Understanding, Listening Effort, Loudness, Loudness Satisfaction, Localization, and Hearing Aid Satisfaction), the data were aggregated by calculating the mean rating of each participant in each condition. Pearson-product moment correlation in the mean rating between the test and retest condition across all participants was then calculated for each variable. ►**Fig. 4** shows that all correlations were significant ($p < 0.01$), with the highest and lowest correlations being Speech Understanding ($r = 0.85$) and Loudness Satisfaction ($r = 0.60$), respectively. The coefficient of repeatability was also calculated for these variables and the results are shown in ►**Fig. 4**. The Hearing Aid Satisfaction variable has the largest (poorest) coefficient of repeatability.

The test-retest correlations of the retrospective questionnaires are shown in ►**Table 4**. All correlations were significant ($p < 0.01$), with the highest and lowest correlations being the Speech subscale of the SSQ ($r = 0.85$) and the

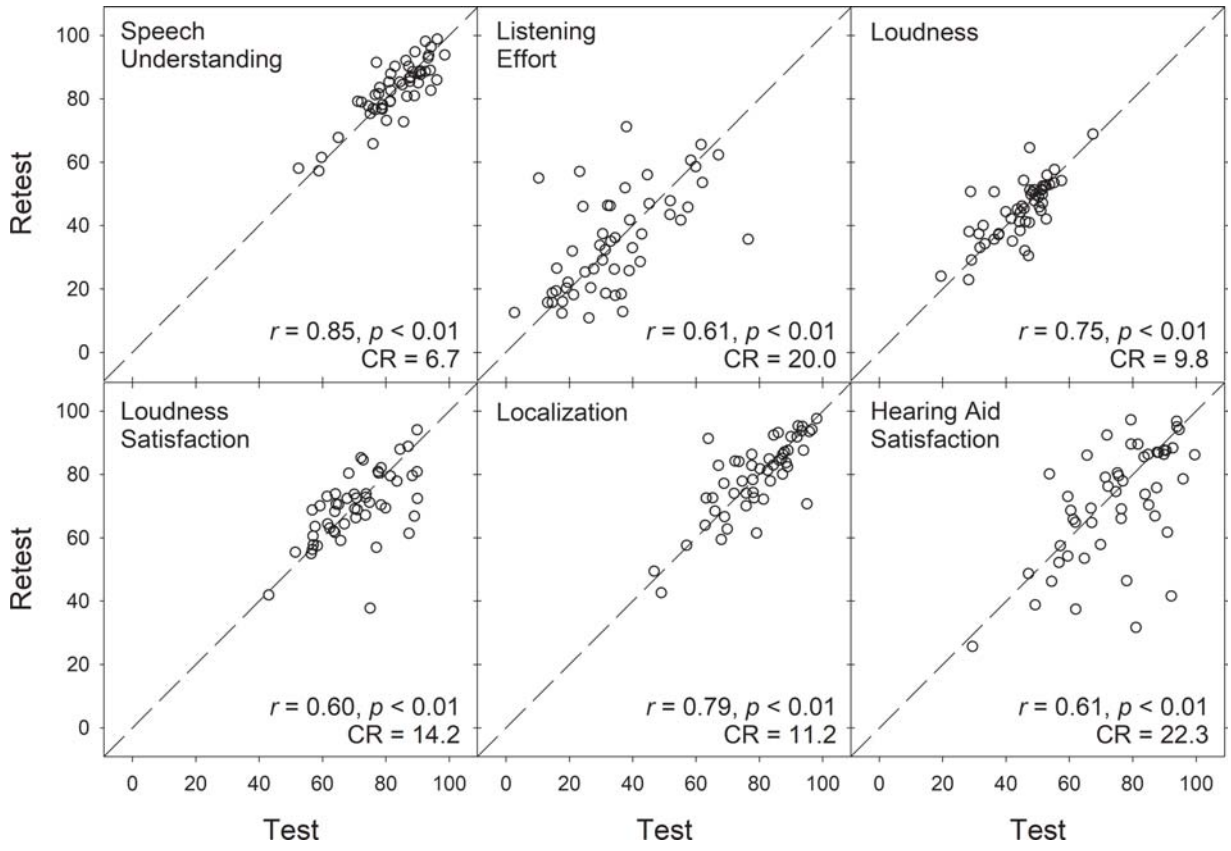

Fig. 4 Scatter plots of the ratings of the interval variables that describe listening experience in the test and retest conditions. Diagonal lines represent perfect match between the test and retest conditions. CR, coefficient of repeatability.

Table 4 Test-retest correlation of four retrospective questionnaires

Questionnaire	Subscale	*r*-Value
HHIE	Emotional	0.76[a]
	Social	0.83[a]
APHAB	Ease of Communication	0.64[a]
	Reverberation	0.74[a]
	Background Noise	0.72[a]
	Aversiveness	0.77[a]
SADL	Positive Effect	0.87[a]
	Service and Cost	0.36[a]
	Negative Features	0.55[a]
	Personal Image	0.63[a]
SSQ	Speech	0.85[a]
	Spatial	0.82[a]
	Qualities	0.82[a]

Abbreviations: AHPAB, Abbreviated Profile of Hearing Aid Benefit; HHIE, Hearing Handicap Inventory for the Elderly; SADL, Satisfaction with Amplification in Daily Life; SSQ, Speech, Spatial, and Qualities hearing scale.
[a]$p < 0.01$.

Service and Cost subscale of the SADL ($r = 0.36$), respectively. The low correlation of the Service and Cost subscale of the SADL could be due to the subscale consisting only of one item in the present study. If this subscale is excluded, the test-retest correlation of the retrospective questionnaires ranged from 0.55 to 0.85.

Discussion

Test-Retest Comparisons

In the present study, older adults with hearing loss used the EMA methodology to report the characteristics of listening contexts and listening experiences in situ in the test and retest conditions. The results first indicated that, at the group level, the characteristics of listening contexts were fairly consistent between the test and retest conditions (►**Table 2**). Although the interaction between response and test-retest condition was significant for the Location variable ($p = 0.02$), the distribution of the probability across the five responses of this variable was still very similar across the test and retest conditions. The similarity in listening context characteristics between the test and retest conditions is in line with Humes et al[29] study that used a hearing aid data-logging sound environment classification feature to objectively quantify the acoustic environments for older adults with hearing loss. Humes et al[29] found that the acoustical environments in which older adults wore hearing aids were quite stable through the first year of usage. The characteristics of many listening contexts reported by the present study were also consistent with the literature. For example, the older participants in the present study reported "Speech listening, media" in approximately 30% of the EMA surveys (►**Table 2**). This is fairly similar to the time spent in television watching reported by Klein et al[30] (~26% of the time), Wu and Bentler [5] (~28% of the time), and Mares and Woodard[31] (3.5–3.8 hours/day). Of interest is that in the present study on average the participants reported "quiet" in the Noisiness question in 58% of the surveys (►**Table 2**). This is close to a group of listeners described by Humes et al[29] who spent almost half of their time in environments classified as quiet (the "Quiet" group). The other two groups identified by Humes et al[29] were the "Speech" group who had the highest proportions of hearing aid usage in environments involving speech and the "Noise" group who had the highest proportions of hearing aid usage in environments involving noise. Because the Quiet group participated in social activities less often compared with the Speech and Noise groups,[29] it is likely that the participants of the present study had relatively inactive social lifestyles.

The results shown in ►**Table 3** further indicated that, at the group level, the self-reported listening experience was similar across the test and retest conditions. This is consistent with the study by Humes et al[32] that used several retrospective self-reports to measure hearing aid outcomes at 7, 15, 30, 60, 90, and 180 days post-fit. Humes et al[32] found that self-reported outcomes were fairly stable over time, especially for outcomes obtained after 15 days post-fit. In the present study, the largest difference in rating between the test and retest conditions was from the Hearing Aid Satisfaction variable (3.9 points out of 100 points), although the difference was not statistically significant. Humes et al[32] found that self-reported hearing aid satisfaction was very stable after 7 days post-fit. One explanation for the relatively large decrease in satisfaction rating is that the lengthy involvement of the larger study (5–6 months) made the participants less enthusiastic in the study and therefore reported lower hearing aid satisfaction in the retest condition. However, this explanation is not supported by the nonsignificant correlation between the degree of satisfaction decrease and the interval between the test and retest conditions ($r = 0.24$, $p = 0.09$).

Test-Retest Reliability

In terms of test-retest reliability, the median correlation coefficients of the categorical variables of the present study (Activity, Location, Talker Location, and Noise Location) ranged from 0.63 to 1.0 (►**Fig. 2**). The Noise Location had the lowest correlation and the largest variation across participants (i.e., the boxplot with the broadest spread in ►**Fig. 2**). This finding likely, at least in part, results from the availability of fewer EMA surveys for analysis. Recall that the Noise Location question was presented only when participants reported that it was not quiet. Since the participants spent most of their time in quiet, the amounts of the surveys containing Noise Location data were 17.1 (test) and 16.1 (retest) per participant, which were less than half of the surveys competed by the participants in each condition (~38 surveys per participant). If the quantity of data on a characteristic in the surveys is small, the aggregated data cannot

offset the variation across individual surveys, resulting in low reliability.

For the ordinal variables (Talker Familiarity, Visual Cues, Noisiness, and Reverberation), the correlation coefficients ranged from 0.46 to 0.89, with the highest and lowest correlations being Reverberation and Talker Familiarity, respectively (►**Fig. 3**). One reason for the low test-retest correlation of the Talker Familiarity variable is that the question itself and its responses were not appropriate for some listening situations. For example, in a group conversation if half of the talkers were familiar to the listener and the other half were strangers, the listener would have difficulty in determining which response to select when answering the talker familiarity question. Other than the Talker Familiarity variable, the test-retest correlations (ranging from 0.64 to 0.89) were fairly similar to those of retrospective questionnaires shown in ►**Table 4** (ranging from 0.55 to 0.85).

Among the six variables that describe listening experience, three had lower test-retest correlations (►**Fig. 4**): Listening Effort (0.61), Loudness Satisfaction (0.60), and Hearing Aid Satisfaction (0.61). The lower reliability of Listening Effort could result from the direction of the visual analog scale of this variable (right side indicating more effortful) being opposite to other EMA questions. Participants might accidentally answer this question in the wrong direction, resulting in less reliable data. Furthermore, the Listening Effort question ("*On average, how much effort was required to listen effectively?*") could be difficult to understand and answer for some participants. Moore and Picou[33] suggested that self-reported listening effort could be biased because people tend to rate their performance, which is easier to answer, instead of rating how much effort it took to achieve the result. This bias could reduce the reliability of self-reported listening effort.

The other two variables that had lower test-retest correlations both involved satisfaction (with loudness and hearing aids). The Hearing Aid Satisfaction had the largest (poorest) coefficient of repeatability (22.3 points) among the interval variables (►**Fig. 4**). This could be due to that the construct of satisfaction is multidimensional, including, for example, positive effects (e.g., reduced communication disability), personal image (e.g., self-image and stigma), negative effects (e.g., feedback problems), and service and cost.[14] Furthermore, satisfaction rating tends to be affected by the context. For example, Faiers and McCarthy[34] compared hearing aid outcomes for those who paid a portion of the cost of hearing aids and those who did not pay. The results indicated that financial outlay of hearing aid users did not affect the reduction in self-reported hearing handicap, while having a significant effect on hearing aid satisfaction (those who paid were more dissatisfied). Because of the multidimensional nature of satisfaction and because the satisfaction level tends to be affected by the context, the variation of satisfaction ratings could be large from moment to moment, resulting in lower test-retest reliability. Regardless, the test-retest correlations shown in ►**Fig. 4** (ranging from 0.60 to 0.85) are fairly similar to those of the retrospective questionnaires shown in ►**Table 4** (ranging from 0.55 to 0.85).

Effect of Number of Surveys

As mentioned earlier, the low test-retest correlation of the Noise Location variable could be due to the availability of fewer EMA surveys for analysis. To explore how EMA survey number could affect test-retest reliability, simulations were conducted on four variables that showed higher test-retest correlations in the present study: Reverberation, Noisiness, Localization, and Loudness. For each variable, the 51 participants provided 1,995 data points (39.4 per person) and 1,905 (37.4 per person) data points in the test and retest conditions, respectively. In the simulation, a certain ratio (e.g., 10%) of the EMA surveys were randomly sampled without replacement in each participant and in each of the test and retest conditions. The sampled data were aggregated for each participant and the test-retest correlation was calculated. The entire process was repeated 1,000 times and the mean and 95% confidence interval of the test-retest correlation were computed. The sampling ratio was varied from 10 to 90% with a 10% increment. The results shown in ►**Fig. 5** indicate that test-retest correlation decreases monotonically as the survey sample size decreases. Based on the figure, it is likely that the reliability is not much affected if the sampling ratio is 50% or higher. Therefore, the simulation suggests that to achieve a reasonable reliability, each participant needs to complete approximately 20 EMA surveys in each condition. Further, the simulation seems to suggest that Localization and Loudness (►**Fig. 5B**) are more resistant to the impact of survey size than Reverberation and Noisiness (►**Fig. 5A**). It is unclear if this difference results from response format (visual analog scale vs. categorical/ordinal scale) or the construct that the survey question is trying to measure (listening experience vs. listening context). More research is needed to systematically determine the effect of survey number on EMA test-retest reliability.

Limitations

The present study has several limitations. First, the test-retest reliability data reported in the present study was specific to the EMA survey shown in ►**Table 1** and may not generalize to surveys that have different numbers of questions, presentation logic, and response formats. For example, an EMA design that contains only few questions with two response options (e.g., like vs. dislike) but has very short interprompt interval (e.g., every 30 minutes) could yield a very different reliability compared with those reported in the present study.

Second, although one of the advantages of the EMA methodology is that it could measure outcomes of an intervention in specific contexts, in the present study the test-retest reliability was computed based on the data aggregated across all kinds of contexts, making the reliability results reported in the present study less useful. Although determining context-specific reliability is desirable, the present study does not have sufficient data points for every context of interest. To illustrate this, ►**Table 5** shows the test-retest reliability of the Hearing Aid Satisfaction variable across all contexts and in each of the four Noisiness categories (quiet, somewhat noisy, noisy, and very noisy). ►**Table 5** also shows the numbers of participants and EMA surveys available for

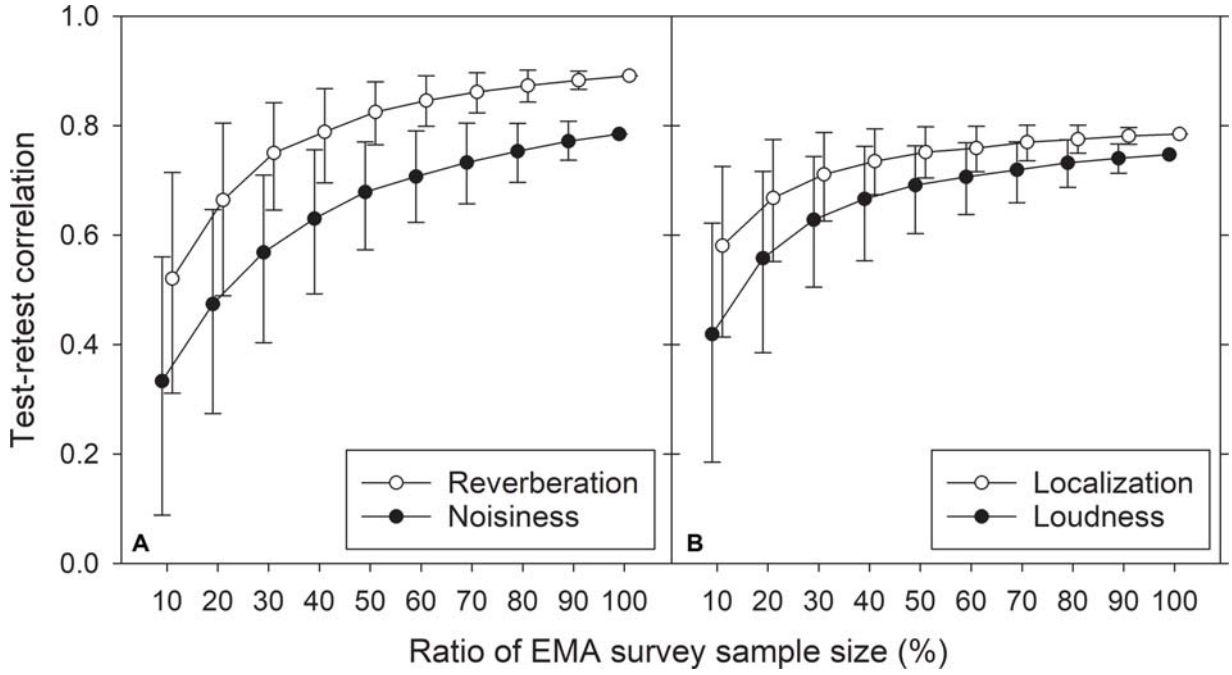

Fig. 5 Impact of survey sample size on test-retest reliability. EMA, ecological momentary assessment.

the analysis. Because the noisy and very noisy categories were rarely reported in surveys (4.6 and 2.3 surveys per person, respectively), few participants (noisy: $n = 25$; very noisy: $n = 14$) had EMA data from both the test and retest conditions. Without sufficient numbers of EMA surveys from each participant, the context-specific reliability in the noisy and very noise categories is likely to be underestimated. Future research that systematically examines context-specific reliability of EMA is needed.

Third, the ordinal variables (Talker Familiarity, Visual Cues, Noisiness, and Reverberation) were treated as interval variables to determine their test-retest reliability. Consecutive integers were assigned to the responses based on the order of response. This approach is arbitrary because there is no empirical evidence to support that the differences between responses (e.g., quiet vs. somewhat noisy and somewhat noisy vs. noisy) are equal across all responses. The reliability the ordinal variables determined in the present study may not reflect the true reliability.

Implications

Although many variables shown in ►**Figs. 2–4** have appropriate test-retest reliability (e.g., Talker Location, Reverberation, and Speech Understanding), the present study suggests that the reliability of EMA could be threatened by many factors. For example, because it is impossible to strictly control real-world environments, EMA typically relies on a large amount of data from each respondent to derive a clear pattern of human experiences and behaviors. If the number of surveys completed by a respondent is not sufficiently large (e.g., the Noise Location in the present study and ►**Fig. 5**), the reliability of EMA could be reduced. Further, like all self-reported questionnaires, a poorly designed question (e.g., the Listening Effort and Talker Familiarity in the present study) could threaten the test-retest reliability. Therefore, the present study stresses the importance of using validated questions and responses in EMA. In most audiology research (including the present study), EMA questions were created specifically for the study and, therefore, their wordings and

Table 5 Test-retest reliability of Hearing Aid Satisfaction across all contexts and in each Noisiness category

		Number of participants	Number of surveys	Test-retest correlation		CR
				r-Value	*p*-Value	
All-context		51	33.5	0.61	< 0.01	22.3
Context-specific	Quiet	48	19.1	0.51	< 0.01	23.4
	Somewhat noisy	49	12.2	0.57	< 0.01	23.1
	Noisy	25	4.6	0.45	0.024	25.6
	Very noisy	14	2.3	0.57	0.033	38.1

Abbreviation: CR, coefficient of repeatability.
Note: The third column shows the numbers of participants who had data in both the test and retest conditions. The fourth column shows the mean numbers of surveys per participant per condition available for reliability analysis.

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. Alternatively, EMA could adapt the questions from validated retrospective questionnaires. For example, the wordings and response formats of the questionnaires Glasgow Hearing Aid Benefit Profile[35] and International Outcome Inventory for Hearing Aids[36] could be used in EMA to assess the respondent's listening experiences in situ. Because paper-and-pencil and electronic administrations of self-reports generally yield equivalent results,[37] it is expected that, with the questions adapted from validated questionnaires, the computer- or smartphone-based EMA would have reasonable test-retest reliability.

Conclusion

EMA in audiology research could have test-retest reliability comparable to validated retrospective questionnaires. To increase reliability, it is important for EMA to have survey questions that are appropriately designed and have sufficient numbers of surveys completed by respondents.

Funding

The present research was supported by NIH/NIDCD R03DC012551 and R01DC015997 and NSF SCH 1838830.

Presentation at Meetings

Portions of this paper were presented at the annual meeting of the American Auditory Society, March, 2017, Scottsdale, Arizona, USA.

Conflicts of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Acknowledgments

None.

References

1 Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1–32
2 Preminger JE, Cunningham DR. Case-study analysis of various field study measures. J Am Acad Audiol 2003;14(01):39–55
3 Walden BE, Surr RK, Cord MT, Dyrlund O. Predicting hearing aid microphone preference in everyday listening. J Am Acad Audiol 2004;15(05):365–396
4 Wu YH, Bentler RA. Impact of visual cues on directional benefit and preference: Part II–field tests. Ear Hear 2010;31(01):35–46
5 Wu YH, Bentler RA. Do older adults have social lifestyles that place fewer demands on hearing? J Am Acad Audiol 2012;23(09):697–711
6 Bentler R, Wu YH, Kettel J, Hurtig R. Digital noise reduction: outcomes from laboratory and field studies. Int J Audiol 2008;47 (08):447–460
7 Galvez G, Turbin MB, Thielman EJ, Istvan JA, Andrews JA, Henry JA. Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users. Ear Hear 2012;33(04): 497–507
8 Timmer BHB, Hickson L, Launer S. Do hearing aids address real-world hearing difficulties for adults with mild hearing impairment? Results from a pilot study using ecological momentary assessment. Trends Hear 2018;22:2331216518783608
9 Wu YH, Stangl E, Chipara O, Hasan SS, DeVries S, Oleson J. Efficacy and effectiveness of advanced hearing aid directional and noise reduction technologies for older adults with mild to moderate hearing loss. Ear Hear 2019;40(04):805–822
10 Wu YH, Stangl E, Zhang X, Bentler RA. Construct validity of the ecological momentary assessment in audiology research. J Am Acad Audiol 2015;26(10):872–884
11 Timmer BHB, Hickson L, Launer S. Ecological momentary assessment: feasibility, construct validity, and future applications. Am J Audiol 2017;26(3S):436–442
12 Weinstein BE, Spitzer JB, Ventry IM. Test-retest reliability of the Hearing Handicap Inventory for the Elderly. Ear Hear 1986;7(05): 295–299
13 Cox RM, Alexander GC. The abbreviated profile of hearing aid benefit. Ear Hear 1995;16(02):176–186
14 Cox RM, Alexander GC. Measuring satisfaction with amplification in daily life: the SADL scale. Ear Hear 1999;20(04):306–320
15 Kramer SE, Goverts ST, Dreschler WA, Boymans M, Festen JM. International Outcome Inventory for Hearing Aids (IOI-HA): results from The Netherlands. Int J Audiol 2002;41(01):36–41
16 Singh G, Kathleen Pichora-Fuller M. Older adults’ performance on the speech, spatial, and qualities of hearing scale (SSQ): test-retest reliability and a comparison of interview and self-administration methods. Int J Audiol 2010;49(10):733–740
17 Kline P. The Handbook of Psychological Testing. 2nd ed. London: Routledge; 2000
18 Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill; 1994
19 Hektner JM, Schmidt JA, Csikszentmihalyi M. Psychometrics of ESM data. In: Experience Sampling Method: Measuring the Quality of Everyday Life. Thousand Oaks, CA: Sage; 2007:103–121
20 Csikszentmihalyi M, Larson R. Validity and reliability of the experience sampling method. In: Csikszentmihalyi M, ed. Flow and the Foundations of Positive Psychology. Dordrecht, The Netherlands: Springer; 2014:35–54
21 Keidser G, Dillon H, Flax M, Ching T, Brewer S. The NAL-NL2 prescription procedure. Audiology Res 2011;1(01):e24
22 Hasan SS, Lai F, Chipara O, Wu YH. AudioSense: Enabling real-time evaluation of hearing aid technology in-situ. Paper presented at the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal2013
23 Stone AA, Broderick JE, Schwartz JE, Shiffman S, Litcher-Kelly L, Calvanese P. Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. Pain 2003;104(1-2):343–351
24 Wu YH, Stangl E, Chipara O, Hasan SS, Welhaven A, Oleson J. Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild to moderate hearing loss. Ear Hear 2018;39(02):293–304
25 Palmer CV, Mueller GH, Moriarty M. Profile of aided loudness: a validation procedure. Hear J 1999;52:34–36
26 Ventry IM, Weinstein BE. The hearing handicap inventory for the elderly: a new tool. Ear Hear 1982;3(03):128–134
27 Gatehouse S, Noble W. The Speech, Spatial and Qualities of Hearing Scale (SSQ). Int J Audiol 2004;43(02):85–99
28 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1 (8476):307–310
29 Humes LE, Rogers SE, Main AK, Kinney DL. The acoustic environments in which older adults wear their hearing aids: insights from datalogging sound environment classification. Am J Audiol 2018; 27(04):594–603
30 Klein KE, Wu YH, Stangl E, Bentler RA. Using a digital language processor to quantify the auditory environment and the effect of hearing aids for adults with hearing loss. J Am Acad Audiol 2018; 29(04):279–291

31 Mares ML, Woodard EH. In search of the older audience: adult age differences in television viewing. J Broadcast Electron Media 2006;50:595–614
32 Humes LE, Halling D, Coughlin M. Reliability and stability of various hearing-aid outcome measures in a group of elderly hearing-aid wearers. J Speech Hear Res 1996;39(05):923–935
33 Moore TM, Picou EM. A potential bias in subjective ratings of mental effort. J Speech Lang Hear Res 2018;61(09):2405–2421
34 Faiers G, McCarthy P. Study explores how paying affects hearing aid users' satisfaction. Hear J 2004;57:25–32
35 Gatehouse S. Glasgow hearing aid benefit profile: derivation and validation of a client-centered outcome measure for hearing aid services. J Am Acad Audiol 1999;10:103
36 Cox RM, Alexander GC. The International Outcome Inventory for Hearing Aids (IOI-HA): psychometric properties of the English version. Int J Audiol 2002;41(01):30–35
37 Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. Value Health 2008;11(02): 322–333