Hearing in Real-Life Environments (HERE): Structure and Reliability of a Questionnaire on Perceived Hearing for Older Adults

Antje Heinrich,[1,2,5] Tuija M. Mikkola,[3,5] Hannele Polku,[4] Timo Törmäkangas,[4] and Anne Viljanen[4]

Objectives: The ability to hear in a variety of social situations and environments is vital for social participation and a high quality of life. One way to assess hearing ability is by means of self-report questionnaire. For questionnaires to be useful, their measurement properties, based on careful validation, have to be known. Only recently has consensus been reached concerning how to perform such validation and been published as COSMIN (consensus-based standards for the selection of health status measurement instruments) guidelines. Here the authors use these guidelines to evaluate the measurement properties of the "Hearing in Real-Life Environments" (HERE) questionnaire, a newly developed self-report measure that assesses speech perception, spatial orientation, and the social-emotional consequences of hearing impairment in older adults. The aim is to illustrate the process of validation and encourage similar examinations of other frequently used questionnaires.

Design: The HERE questionnaire includes 15 items with a numeric rating scale from 0 to 10 for each item and allows the assessment of hearing with and without hearing aids. The evaluation was performed in two cohorts of community-dwelling older adults from Finland (n = 581, mean 82 years) and the United Kingdom (n = 50, mean 69 years). The internal structure of the questionnaire and its relationship to age, hearing level, and self-reported and behavioral measures of speech perception was assessed and, when possible, compared between cohorts.

Results: The results of the factor analysis showed that the HERE's internal structure was similar across cohorts. In both cohorts, the factor analysis showed a satisfactory solution for three factors (speech hearing, spatial hearing, and socio-emotional consequences), with a high internal consistency for each factor (Cronbach's α's for the factors from 0.90 to 0.97). Test–retest analysis showed the HERE overall mean score to be stable and highly replicable over time (intraclass correlation coefficient = 0.86, standard error of measurement of the test score = 0.92). The HERE overall mean score correlated highly with another self-report measure of speech perception, the Speech Spatial Qualities of Hearing questionnaire (standardized regression coefficient [β] = –0.75, $p < 0.001$), moderately highly with behaviorally assessed hearing level (best-ear average: β = 0.45 to 0.46), and moderately highly with behaviorally measured intelligibility of sentences in noise (β = –0.50, $p < 0.001$).

Conclusions: Using the COSMIN guidelines, the authors show that the HERE is a valid, reliable, and stable questionnaire for the assessment of self-reported speech perception, sound localization, and the socio-emotional consequences of hearing impairment in the context of social functioning. The authors also show that cross-cultural data collected using different data collection strategies can be combined with a range of statistical methods to validate a questionnaire.

Key words: Aging, COSMIN criteria, Hearing, Questionnaire validation, Speech perception.

(Ear & Hearing 2019;40;368–380)

INTRODUCTION

Accurate speech perception, which is vital for successful communication, good social participation, and a high quality of life (Cruice et al. 2006), depends partly on good hearing sensitivity. Hearing can be assessed in a number of ways, with different methods preferred by different fields. One method, preferred by experimental studies and clinical settings, is the computation of pure-tone averages (PTA), a combined sensitivity threshold of basic tone stimuli at a number of frequencies. This objective measure correlates well with speech perception of simple stimuli by listeners with hearing loss in quiet listening situations (Era et al. 1986; Helfer & Wilber 1990; Humes & Roberts 1990; van Rooij & Plomp 1990, 1992; Humes & Christopherson 1991; Jerger et al. 1991; Humes et al. 1994; Divenyi & Haupt 1997; Jerger & Chmiel 1997). But it correlates less well when listeners have good hearing and the listening situation is complex (Duquesnoy 1983; van Rooij et al. 1989; Jerger et al. 1991; Besser et al. 2012; Heinrich et al. 2015, 2016b). Moreover, obtaining these sensitivity measurements requires one-to-one testing in a quiet environment using specialist equipment.

Another way to assess hearing is to use self-report questionnaires. This subjective measure is commonly used in hearing aid validation (Whitmer et al. 2015) and in large population-based epidemiological studies (Kramer et al. 2002; Viljanen et al. 2014; Mikkola et al. 2016). In contrast to sensitivity measurements, questionnaire assessment does not require a specialized environment nor one-to-one testing. Moreover, questionnaires assess functioning directly for a variety of real-life situations and environments (for a discussion of the term functioning within the realm of hearing impairment, see Heinrich et al. 2016a) rather than measuring sensitivity of basic stimuli from which functioning is then inferred. Finally, they not only assess hearing functioning but also socio-emotional and other consequences of perceived difficulties. Over the years, a large number of hearing questionnaires have been developed (Whitmer et al. 2015). The main disadvantage of questionnaires is their often insufficient validation, possibly due to a lack of consensus over validation requirements. This consensus has recently

[1]Medical Research Council Institute of Hearing Research, School of Medicine, The University of Nottingham, United Kingdom; [2]Manchester Centre for Audiology and Deafness, School of Health Sciences (ManCAD), University of Manchester, United Kingdom; [3]Folkhälsan Research Center, Helsinki, Finland; and [4]Gerontology Research Center, Faculty of Sport and Health Sciences, University of Jyväskylä, Finland; [5]These Authors contributed equally to this work.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

been reached with the publication of the COSMIN (consensus-based standards for the selection of health status measurement instruments) checklist (Mokkink et al. 2010). This article demonstrates how COSMIN can be used to guide the validation of a hearing-related self-report questionnaire.

Development of the "Hearing In Real-Life Environments" Questionnaire

One, maybe the easiest option, would have been to validate one of the existing hearing questionnaires such as the Speech Spatial and Qualities of Hearing scale (SSQ; Gatehouse & Noble 2004), the Hearing Handicap Inventory for the Elderly (HHIE; Ventry & Weinstein 1982), or the Abbreviated Profile of Hearing Aid Benefit (APHAB; Cox & Alexander 1995). However, we wanted to validate the questionnaire within the context of a large project that investigated the relationship between physical mobility, social participation, and quality of life among older adults (Rantanen et al. 2012). To do this, we sought a questionnaire that (1) would measure functioning of hearing and speech perception in a variety of social situations and environments relevant for older adults, (2) would measure the socio-emotional consequences of hearing difficulties in these situations, (3) would be quick to complete, (4) could be equally used for listeners with and without hearing aids, and (5) would generate data usable in parametric analyses, that is have a fairly large range of response categories. The table in Supplement Digital Content 1, http://links.lww.com/EANDH/A463, compares some of the questionnaires popular within the fields of audiology and hearing science with regard to these requirements and shows that none of the existing questionnaires fulfilled all our requirements.

Consequently, we developed a 16-question self-report instrument, which assesses, on an 11-point scale (0 to 10), perceived functioning in everyday communication situations, localization abilities important for public spaces, and socio-emotional consequences of perceived functional limitations of hearing and that could be used for listeners with and without hearing aids. The questions were loosely based on a collection of items from the APHAB, SSQ, and HHIE but were adapted in three important ways: (1) some questions were modified to better reflect the respondents' cultural reality; (2) all questions were translated into Finnish; and (3) the range of response categories was increased. The translation into Finnish was completed by the native Finnish-speaking authors. The questionnaire was called "Hearing in Real-Life Environments" (HERE).

Although development and translation of the HERE questionnaire predated the recent publication of guidelines for translating and adapting hearing-related questionnaires for different languages and cultures (Hall et al. 2018) by several years, we are pleased that our translation and adaptation process conformed to the guidelines. This was made easier by the fact that the differences in education, literacy, and culture were limited between the populations on which questionnaires were based that inspired some of the questions (US, UK) and the target population for the HERE (Finnish). At the initial English-to-Finnish translation and adaptation during the questionnaire development stage, several English-Finnish bilingual (L1 Finnish) translators with a variety of expertise were involved.

Following its conception, the first step of validation was the examination of its intended structure. This was done by Polku et al. (2018) using a population-based sample of community-dwelling older adults with and without hearing aids. An exploratory factor analysis (EFA) revealed a three-factor structure with an internal consistency (kappa coefficient) for all three subscales of 0.89 to 0.96 (Polku et al. 2018). One of the original 16 questions ("I find traffic noises uncomfortably loud") showed weak loadings on all three factors and was subsequently dropped.

Following this first assessment, a more thorough validation of the questionnaire and in-depth assessment of its psychometric characteristics was necessary to make the questionnaire useful for a larger audience. The details of this validation process are the topic of this article. We based our assessment on the COSMIN guidelines (Mokkink et al. 2010) because they give clear recommendations for how to evaluate the measurement properties of health-related patient-reported outcomes by specifying six categories (Points) of assessment: (1) structural validity; (2) internal consistency; (3) test–retest reliability; (4) measurement error; (5) hypothesis testing; (6) cross-cultural validity. Questionnaire evaluation was carried out using two samples. First, the Finnish sample (subsequently called LISPE) was the same as that used by Polku et al. (2018) but limited to non-hearing aid wearers. The sample was further limited in some analyses by only including those non-hearing aid wearers whose hearing sensitivity thresholds were known (LISPE2). The second sample was a sample collected in the United Kingdom (subsequently called UK sample). Again, differences in education and literacy between source (Finland) and target (UK) populations were limited, and thus translation was fairly straightforward. In the back-translation into English, both Finnish-English bilingual speakers and English speakers with clinical and research expertise were included and checked for adequacy of translation and concepts.

Whenever possible, parallel analyses were conducted in both samples and results compared between them. Table 1 details which analysis methods were used to investigate each aspect of methodological quality and which sample was used.

In a first step, we reexamined the factor structure found by Polku et al. (2018) because we could not assume that the structure would remain unchanged when hearing aid wearers were removed from the sample. We then compared the questionnaire structure obtained for the Finnish sample to that of the UK sample, which was smaller and which differed in hearing loss, social participation, and cultural background. Doing this allowed us to assess to what extent the questionnaire shows a comparable internal structure across participant groups in general (Point 1) and across cultures in particular (Point 6). One advantage of the smaller UK sample was that we could collect additional self-report and behavioral measures. This in turn allowed us to investigate the construct validity of the HERE questionnaire by understanding how a particular questionnaire relates to other measures that are commonly used in the field (Point 5) such as PTA (calculated here as best-ear average, BEA), the score from the SSQ speech perception subscale (SSQ_{speech}), and behavioral estimates of speech perception in noise. Using self-report and behavioral measures as comparison data helped us to avoid undue influence from method variance (Campbell & Fiske 1959). We either kept the assessment method between measures identical by comparing HERE questionnaire with the SSQ_{speech} or kept the complexity of the listening situation similar by comparing the HERE with behaviorally assessed speech intelligibility of sentences in noise.

TABLE 1. Datasets and analysis approaches used to investigate each COSMIN (Consensus-based Standards for the Selection of Health Measurement Instruments) criterion

COSMIN Criteria	Dataset Used	Analysis Approach
1. Structural validity	LISPE and UK	Exploratory factor analysis
2. Internal consistency	LISPE and UK	Cronbach's α
3. Test–retest reliability	UK	Intraclass correlation coefficient (ICC)
4. Measurement error	UK	Standard error of measurement (SEM)
5. Hypothesis testing: relationships with other measures that are commonly used in the field	LISPE2 and UK: relationships with age and BEA UK: relationship with behavioral speech perception in noise and self-reported speech perception (SSQ_{speech})	Linear regression analysis Pearson product–moment correlation
6. Cross-cultural validity	LISPE, LISPE2, and UK	Comparison across samples of different cultures

BEA, best-ear average; LISPE, Life-Space Mobility in Older Age study; LISPE2, subsample of LISPE with available BEA data; SSQ, Speech Spatial Qualities of Hearing scale; UK, United Kingdom sample.

MATERIALS AND METHODS

LISPE Sample

Participants • The sample was recruited as part of the "Life-space mobility in old Age" (LISPE) project, a Finnish population-based 2-year prospective cohort study of community-dwelling older adults for which study design and methods have been reported elsewhere (Rantanen et al. 2012). The present analysis uses cross-sectional data gathered in the second follow-up. Briefly, at baseline, a random sample of 2550 older community-dwelling persons between 75 and 90 years of age was drawn from the Finnish national population register. Of those, 2269 were contacted by letter or telephone to check willingness and eligibility to participate based on the following inclusion criteria: (1) community-dwelling in the study area, and (2) able to understand questions and provide clear answers. Hearing status was not a participation criterion. As a result, a total of 848 older adults were included in the first stage of the study and participated in the baseline interviews, conducted in their homes in the first half of 2012. The follow-up consisted of a telephone interview and a postal questionnaire, conducted 2 years after the baseline measurements, in 2014. Seven hundred sixty-one people participated in this follow-up. Of those, 712 participants returned the postal questionnaire, which included the HERE and questions on quality of life, mood, and mobility. They form the analytic sample in this study (termed LISPE sample hereafter). Those who declined tended to be older, more often lived with a spouse or others, more often perceived their health as poor or very poor, perceived more difficulties with outdoor mobility, and moved outdoors less often than those who participated in the study. While no information on their hearing was available, it is possible that they also had poorer hearing on average than those who participated. For more detailed information on participant attrition in this and the following sample, see Rantanen et al. (2012) and Polku et al. (2018).

In addition, a random sample of 230 older adults was drawn from the original LISPE baseline cohort in January 2014 and screened for eligibility and willingness to participate in assessments of physical performance, cognition, and sensory functions using the same inclusion criteria as applied to the original sample. A total of 169 people (of whom 161 also returned the questionnaire) agreed to participate in the substudy, and audiometric measurements were conducted for 168 (termed LISPE2 hereafter) in their homes during the spring of 2014. Those who did not agree to participate (n = 34) were of the same age (82.6 versus 82.7 years) and had similar sex distribution (65% versus 63% women) to those who participated. A subgroup of those who had declined participation but had answered the HERE (n = 25) had similar self-reported hearing according to Question 1 of the HERE (median 2 versus 3, $p = 0.264$) to those who agreed to participate in LISPE2. Hence, while participants in the LISPE sample are likely to be better functioning on average than the whole target population, participants of LISPE2 are likely to be a good representation of the LISPE population.

Thus, while the original study sample supplied questionnaire data only, the second, smaller sample supplied audiometric data in addition to the questionnaire data they had supplied as part of Finnish LISPE sample. Because the data of the Finnish sample were to be compared with a UK sample of participants explicitly selected to exclude hearing aid users, only data from Finnish participants who indicated that they did not own a hearing aid will be considered in this article. This further reduced the sample size in the whole LISPE sample to n = 581 and in LISPE2 to n = 129. Sample characteristics including age and gender of the whole LISPE sample and the LISPE2 subsample are given in Table 2. The age difference between the overall LISPE sample and its subsample (LISPE2) was not significant.

The LISPE project and its substudy were approved by the Ethical Committee of the University of Jyväskylä. Participants were informed about the project and signed a written consent form.

Materials

HERE Questionnaire • Self-reported hearing ability was assessed using the HERE questionnaire. The questionnaire assesses perceived difficulty and resulting socio-emotional consequences of communicating in situations potentially relevant for older adults. The questionnaire is designed to measure perceived functioning for three aspects of listening: (1) listening to speech in a variety of situations and environments, (2) locating sound in space, (3) the consequences of hearing impairment for emotional well-being and social participation. Originally, the questionnaire contained 16 questions; however, one ("I find traffic noises uncomfortably loud") was dropped after Polku et al's (2018) original analysis. All questions were answered by choosing a number between 0 and 10 corresponding to the participant's perceived hearing ability in the particular

TABLE 2. Demographic information on gender and age of the LISPE, LISPE2, and UK samples

		Gender (M/F)	Age			BEA		
	n	(%)	M	SD	Range	M	SD	Range
LISPE sample	581	37/63	82	4.0	76–91	NA		
LISPE2 sample	129	36/64	82	4.2	76–91	39	11	16–64
UK sample	50	44/56	69	6.4	61–86	20	9.9	5–46

Also given are the best-ear averages (BEA) when pure-tone audiometry was available.
LISPE, Life-Space Mobility in Old Age; M, mean; M/F, male/female; n, number of participants; SD, standard deviation.

situation. Higher scores indicated poorer performance or more difficulty. Numbers were presented along a continuum below each question. Additionally, above the extremes (0 and 10), a verbal description was provided (e.g., 0 = no difficulty at all, 10 = very difficult). Given that hearing status was not an exclusion criterion to participate in the original study (Rantanen et al. 2012), the questionnaire needed to be applicable to both hearing aid wearers and non-hearing aid wearers. For participants who used a hearing aid, two separate answers were required for each item: without hearing aid (A) and with hearing aid (B), following methodology originally developed for the APHAB (Cox & Alexander 1995).

One item of the socio-emotional scale, Q8 "I need help from other people because of my hearing difficulty," was erroneously not translated into English and was therefore missing from the English version of the questionnaire. To account for this difference, all analyses of LISPE and LISPE2 excluded Q8. The absence of the question is indicated by the retention of the original numbering. Hence, after accounting for the missing question, 14 questions from the HERE questionnaire were included in all further analyses.

Procedure

Pure-Tone Audiometry • Hearing was assessed only for the smaller random sub-sample (LISPE2). Pure-tone air conduction thresholds were measured at octave frequencies between 0.125 and 8 kHz in the participants' homes using pure-tone screening audiometry (Oscilla USB-330, Inmedico A/S, Denmark) and Peltor noise reducing headphones with a noise reduction rating of 21 dB SPL. Both ears were measured separately. Hearing thresholds were estimated using the automatic Hughson-Westlake protocol in which two out of three correct answers determined the lowest sound intensity the subject is able to hear. The maximum sound intensity was 90 dB SPL. If the participant could not hear this intensity, 100 dB SPL was recorded as the hearing threshold. The audiometry data were automatically stored on a personal computer. PTAs of the octave frequencies between 0.25 and 8 kHz were calculated separately for each ear as an arithmetic mean over all measured frequencies. The mean interaural difference between the two ears was 7.6 dB (range: 0 to 45; SD = 8.0). When hearing thresholds were included as part of subsequent analyses, the lower average of the two ears, the BEA, was used. Its group statistics are reported in Table 2.

HERE Questionnaire • The questionnaire, plus instructions about how to complete it, was posted to participants of the Finnish study as part of a larger selection of questionnaires.

UK Sample

Participants • The UK sample consisted of n = 50 older community-dwelling adults who had responded to an advertisement for a study investigating age-related changes in speech-in-noise perception. The data were part of a bigger experimental study that investigated the contributions of auditory and cognitive factors to behavioral and self-reported aspects of speech-in-noise perception. Only behavioral and self-reported aspects of speech perception and hearing sensitivity will be discussed here. None of the unreported results relate to the topics discussed in this article. Potential participants were screened for hearing, language status, and neurological function, and the following inclusion criteria were used: (1) over 60 years of age; (2) native English speakers; (3) absence of neurological disorders and psychoactive medication; and (4) no hearing aid use. Participant characteristics of the sample are reported in Table 2. Participants came in for two testing sessions, typically about 1 week apart, in which auditory and cognitive functioning as well as speech perception were measured in a number of tasks. The HERE questionnaire was usually completed in the first testing session as part of a number of questionnaires, which also included the speech section of the SSQ. Pure-tone audiometry was also completed during the first visit. Speech perception measures were acquired during both visits. Ethical approval was obtained from the University of Nottingham School of Psychology (Ref 464). All participants were informed about the study, signed a consent form, and were paid an inconvenience allowance of £7.50/hour.

The age difference between the UK sample and LISPE samples was significant ($t(52.3) = 13.4$ (equal variances not assumed), $p < 0.001$), with participants in the UK sample being significantly younger than their Finnish counterparts. The gender distribution between the samples did not systematically vary ($\chi^2 = 0.87$, $p = 0.352$).

Materials

HERE Questionnaire • As the questionnaire was originally developed for use in a Finnish study, its questions were translated into English for use in the UK study by a team of Finnish and native English speakers, most of whom were part of the HEARATTN consortium (Heinrich et al. 2016a).

SSQ_{speech} Questionnaire • The Speech section of the SSQ questionnaire (Gatehouse & Noble 2004) was administered. It contains a section of 14 questions dedicated to speech perception in a variety of situations. These situations include speech perception in conversational and nonconversational settings (radio/TV), the use of visual information, and listening situations that are similar in their conversational setting but differ in acoustic parameters (one or several background talkers, voices with similar or different pitches). This is in contrast to the HERE whose speech perception section only comprises seven questions and only enquires about conversational settings. Nevertheless, given that both questionnaires assess aspects of speech perception, we

assume that the scales are similar enough to expect a significant correlation between scores within the same person.

Speech Perception • We measured speech perception behaviorally by asking UK participants to listen to and repeat short sentences presented in background noise. Accuracy of detection of the final word of each sentence was the variable of interest. Stimuli were 112 sentences from a recently developed sentence test (Heinrich et al., unpublished). This test, chiefly developed to test the effect of semantic predictability on speech perception, contained sentence pairs with identical final (target) words but differing preceding sentence bases. The sentence bases were chosen in such a way that the (identical) final word in one sentence was highly predictable from the preceding context while in the other sentence of the pair was less predictable (e.g., "We'll never get there at this rate" versus "He's always had it at this rate"). Only one sentence of each pair, either the high- or low-predictability version, was heard by a single participant. Half the sentences heard by each participant were the less predictable sentence of each pair and half were the more predictable sentence. All sentences were presented in speech-modulated noise derived from the input spectrum of the sentences. Half of all sentences were presented at a signal to noise ratio (SNR) of −4 dB and half at an SNR of −7 dB.

Procedure

Pure-Tone Audiometry • In the UK sample, pure-tone air conduction thresholds were measured across the same range of frequencies as was used in the LISPE2 sample, namely, octave frequencies between 0.125 and 8 kHz. Measurements were taken as part of the first of two visits to the hearing laboratory at the Medical Research Council Institute of Hearing Research in Nottingham. Testing was carried out in a double-wall sound-attenuating booth (Industrial Acoustics Company, Winchester, UK) using an Interacoustics Audiometer AT235 (Interacoustics, Middelfart, Denmark) and TDH39P headphones (Telephonics, Farmingdale, NY). Both ears were measured separately; the Hughson-Westlake protocol was implemented manually and followed the recommendations of the British Society of Audiology (2011). Again, the maximum sound intensity was 90 dB SPL, and if the participant could not hear this intensity, 100 dB SPL was recorded as the hearing threshold. The PTAs were calculated separately for each ear as an arithmetic mean over 0.25 and 8 kHz. On average, interaural differences were 4.5 dB in this group (range: 0 to 29; SD = 4.9). The lower of the two PTAs per ear is reported as BEA in Table 2. The difference in BEA between the LISPE2 and UK samples was significant (t (177) = 10.6, $p < 0.001$), with the UK sample having lower thresholds on average.

HERE Questionnaire • Participants filled in the HERE questionnaire twice: initially as part of a larger number of questionnaires during the first of two behavioral testing sessions, and again when asked to repeat the questionnaire either during laboratory visits in the context of other studies or as part of a postal follow-up. The mean time interval between the test and retest was 158 (SD = 94) days.

SSQ_{speech} • The speech section of the SSQ (first 14 questions) was filled in as part of a larger number of questionnaires during the first of two behavioral testing sessions.

Speech Perception • Testing was carried out in the same double-walled sound-attenuated chamber that was used for the audiometric testing, but using Sennheiser HD280 headphones. All testing was in the left ear only. As the study consisted of two visits around 1 week apart, the speech perception task was tested in the first session for half of the participants and in the second session for the other half. The overall presentation level of the target stimuli was individually adjusted to be 30 dB above each listener's speech reception threshold (dB sensation level). This was done to partially account for differences in hearing level. Sentences were presented in blocks of high/low predictability and high/low SNR, with block presentation counterbalanced across participants. The 13 sentences within each block were randomized. For the purpose of the current study, the intelligibility scores for both types of sentences and both types of SNR were averaged to create a single score. Testing was self-paced. Speech perception scores were "rationalized" arcsine (RAU)-transformed (Studebaker 1985) to linearize the s-shaped psychometric function of normal speech perception.

Data Analysis

Mplus version 7 was used for the exploratory factor and regression analyses. (Muthén & Muthén 1998–2010). IBM SPSS Statistics 22 (SPSS Inc. 2013) was used for computing Cronbach's α, intraclass correlation coefficients (ICC), standard error of measurement (SEM), and various Pearson product–moment correlations. In all cases, descriptive results of each sample are presented first and comparisons between samples second. As the main intention of the study was to validate the HERE questionnaire following the COSMIN criteria, the criterion assessed by a particular analysis and the sample on which the analysis was performed are given in brackets (see also Table 1).

Questionnaire Structure (COSMIN Points 1, 2, and 6; LISPE and UK Samples) • We used censoring as a technique to model variables with floor effects. In a first step we explored the number of non-noise factors using Cattell screen plots and parallel analysis. Parallel analysis is a simulation technique to determine non-noise factor number (Horn 1965), which we had adapted for censored variables. Both of these analyses indicated that there were three non-noise eigenvalues in the LISPE data and one in the UK data. Because of large structural similarities across the two data sets, we nevertheless extracted three factors for both localities to enable cross-locality comparisons. Thus, for the UK data, some overfitting may have occurred.

We then analyzed the structure of the hearing questionnaire using an exploratory factor analysis (EFA) appropriate for multivariate censored data and likelihood-based inference (Kamakura & Wedel 2001) and an oblique GEOMIN rotation of the solution (Muthén & Muthén 1998–2010). EFA is a commonly employed technique to test the construct validity of a questionnaire (Bolarinwa, 2015). Using an oblique rotation allows for correlations between factors. Missing data were assumed to have been generated by the missing-at-random mechanism. This missing data mechanism is accounted for by the construction of the fitting function for the weighted least square parameter estimator (WLSMV) estimator (to be used in the regression analyses) in the Mplus software. More detailed information on the WLSMV estimator can be found in Supplement Digital Content 2, http://links.lww.com/EANDH/A464. Estimating factors using this modeling strategy enabled us to include partial response patterns and to only exclude respondents from the analysis when all item scores were missing. As scree plots and parallel analysis indicated three non-noise factors, a three-factor solution was requested in the EFA.

Following best practice guidelines for EFA, two measures of fit for the derived latent structure were calculated: (1) communality (Costello & Osborne 2005) and (2) Cronbach's α (Mokkink et al. 2010). Communality, calculated by subtracting the item residual variance from the observed item variance, indicates the amount of variance accounted for in each item by the three factors. Communality pertained to the questionnaire items and spoke to the structural validity of the questionnaire (COSMIN Point 1). Cronbach's α measured the internal consistency of the items for the derived latent factor and spoke to the question of internal validity (COSMIN Point 2). Additionally, the factor structures obtained for the LISPE and UK samples were compared qualitatively as a way to check for cross-cultural validity and stability of questionnaire structure across samples (COSMIN Point 6).

Two types of factor scores were calculated for each participant. First, individually weighted (also known as factor-weighted) scores were calculated by summing the products of respective factor loadings and all three standardized observed item scores. This technique weighs the standardized item scores by the respective factor loading, thus giving more weight to items that load higher on the factor. Due to the nature of the EFA, these loadings are specific to the sample from which the factor structure is derived and may not transfer to other samples. Therefore, we also calculated unitary-weighted factors as an alternative measure. For this type of scoring, the particular factor structure of the sample is immaterial as all observed scores on a factor receive an identical weight, leading to an equally weighted average of questions for each factor. In contrast to the individually weighted scoring scheme, where the loadings on all three factors are taken into account for each question score, for unitary-weighted mean scores, an item is only included in the factor for which the item had the strongest loading. (Note, however, that this simplified loading structure does not remove any measurement error from the observed scores.) Using unitary loadings can greatly simplify the understanding of the factor analysis because as long as the general factor structure of the questionnaire is replicable across samples, small differences in factor loadings between samples are inconsequential (Floyd & Widaman 1995; Akeroyd et al. 2014). Therefore, unitary-weighted scores were used in regression analyses. In all regression analyses, BEA and the unitary-weighted scores for one of the three latent factors were entered simultaneously into the model.

Test Validity and Reliability (COSMIN Points 3 and 4; UK Sample) • For estimation of test–retest reliability of the HERE in the UK sample, we calculated ICC for each of the three factors and for the total score following the procedure by Weir (2005). To decide which type of ICC was appropriate for the current data set, we first tested whether a significant difference existed between the first and second assessment round by conducting a series of paired t tests for each factor and the overall score. As no difference was detected in any of the t tests (mean differences from −0.19 to 0.16, p from 0.272 to 0.626), we used $ICC_{3,1}$ that assumes (1) no systematic differences or error (consistency), (2) participant scores are available for all rounds (two-way model), (3) a fixed model (no attempt to generalize the results beyond the confines of the study), and (4) a single score for each subject for each round (Weir 2005). The formula is as follows:

$$ICC_{3,1} = \frac{MS_S - MS_E}{MS_S + (k-1)MS_E}$$

where MS refers to mean square, subscript S refers to between-subjects error and subscript E to within-subject error. The mean squares were obtained from analysis of variance tables using SPSS.

We also calculated the SEM and minimal difference (MD) for the test–retest data. SEM indicates the absolute magnitude of trial-to-trial noise in the data, that is, the "typical error" that is expected for an individual (Hopkins 2000; Weir 2005). While ICC is unitless, SEM has the same units as the measurement. SEM is also considered to be independent of the study sample (i.e., not affected by between-subjects variability; Nunnally & Bernstein 1994). The SEM is calculated as follows:

$$SEM_s = SD\sqrt{1 - ICC}$$

where SD is the standard deviation of the scores from all participants and SEM_s is the standard error in estimating the observed scores. However, more interesting is the question of within which boundaries the true scores lie. The SEM_{ts} as the basis for 95% confidence intervals provides this information and is calculated as

$$SEM_{ts} = SD\sqrt{ICC(1 - ICC)}$$

Finally, we wanted to define the smallest difference/change in the scale that can be considered as a real difference (with 95% confidence) in individual scores. Such minimal difference is calculated as

$$MD = SEM \times 1.96 \times \sqrt{2}$$

and indicates the minimum amount by which the scores needs to change for a change to be considered real. For a more detailed discussion of ICC, SEM, and MD, see Weir (2005).

Association Between Questionnaire and Other Measures (COSMIN Point 5; LISPE2 and UK Samples) • We assessed the relationship of the HERE to other measures by relating questionnaire scores to BEA, the speech scale of the SSQ, and a number of speech-in-noise measures.

The relationship with BEA was explored in a series of regression analyses, with perceived hearing as the dependent variable and measured hearing sensitivity and age as independent predictors. Left censoring of the items was taken into account by the "censored below" option and regression coefficients were estimated using the WLSMV (Muthén et al., Reference Note 1). Regression analyses were run separately for LISPE2 and UK samples, and their regression coefficients were compared using the rescaled difference chi-square test of Satorra (2000). This procedure compares a model in which the regression models are fixed as equal between the samples (constrained) to a model in which the regression parameters are free (unconstrained). Note that this test is similar to the conventional likelihood ratio test for testing model constraints, with the exception that it enables such tests to be conducted on, for example, censored variables. A chi-square p value <0.05 indicates that the constrained regression model fits the data significantly worse than the unconstrained model. Adjustment for p values was made by the false discovery rate correction function (Benjamini & Hochberg 1995) implemented in the base package (stats) of the

R programming environment (version 3.4.0). Note that the corrected *p* values are also called *q* values.

In addition to the regression analyses, in the UK sample, Pearson product–moment correlations were calculated for unitary-weighted HERE scores and unitary-weighted SSQ_{speech} scores and for unitary-weighted HERE scores and speech perception scores.

RESULTS

Questionnaire Scores

LISPE Sample • Although responding to all questions was encouraged, questionnaires were not discarded for missing responses to single questions. Missing response rates were between 0.3% (n = 2, Q1) and 5.5% (n = 32, Q3). The mean scores of the 14 questions varied between 0.9 and 3.1 points. In most questions (9), the entire range of the response scale (0 to 10) was used.

UK Sample • No data were missing. The mean scores of the questions varied between 1.3 and 3.5, but participants tended not to use the full range of potential responses to answer questions: for six questions, the response range was 0 to 7, for another five it was 0 to 8. The remaining three questions used a range of 0-9.

A question-by-question comparison between the LISPE and UK samples is provided in Supplement Digital Content 3, http://links.lww.com/EANDH/A465, which shows largely comparable results, particularly when *p* values were false discovery rate adjusted to account for multiple comparisons. In fact, only the group difference in Q11 scores remained significant after the adjustment. SD appeared to be slightly smaller for Q1 to Q6 and Q9 to Q11 in the UK sample, indicating a greater homogeneity of responses.

Exploratory Factor Analysis (COSMIN Point 1 and 2; LISPE and UK Samples)

Factor Structure

LISPE Sample • An EFA with a pre-set three-factor structure on the results of the LISPE sample resulted in Q1 to Q7 loading highest on Factor 1, Q9 to Q11 on Factor 3, and Q12 to Q15 on Factor 2. The communality, which indicates the proportion of variance accounted for by the three factors in each item, varied between 0.67 (Q1) and 0.87 (Q14). These results are all well above the suggested 0.4 cutoff of acceptable communality (Costello & Osborne 2005). Loadings of each item on each factor as well as communality scores for each item for the LISPE sample are presented in Supplement Digital Content 4, http://links.lww.com/EANDH/A466. The same information for the UK sample is presented in Supplement Digital Content 5, http://links.lww.com/EANDH/A467. A visual comparison between loadings for each question is presented in Figure 1.

UK Sample • The EFA showed that Q1 to Q7 loaded highly on Factor 1, Q9 to Q11 on Factor 3, and Q12 to Q15 on Factor 2, with communality varying between 0.55 (Q2) and 0.98 (Q13), again well above the cutoff for acceptable communality.

The biggest difference between the EFA solutions for the two samples was that in the LISPE sample, each questionnaire item loaded strongly only on one of the three factors. In contrast, in the UK sample, all but one questionnaire item (Q2) in Factor 1 also showed a substantial secondary loading on another factor. In the case of Q3 to Q7, this secondary loading was on Factor 2, while for Q1 it was on Factor 3. None of the items primarily loading on Factors 2 or 3 had significant secondary loadings, thereby mirroring the LISPE results. The factors differ in how well they represent the results of the two samples. This is illustrated in Figure 1, which shows that the residual variances of Q2 to Q6 for Factor 1 are lower in the LISPE than in the UK sample, indicating that the latent Factor 1 explained a larger proportion of item variance in the LISPE than in the UK sample for these items. For Q1, Q7, and Q11 to Q13, the opposite was true. For the remaining items, residual variances were roughly similar. Despite these differences and the greater range in communality values, mean communalities and thus the overall fit of the factor structure were very similar between the two samples (LISPE, 0.79; UK sample, 0.79). In concordance with Polku et al. (2018), the latent factors were called "speech hearing", "spatial hearing," and "socio-emotional consequences".

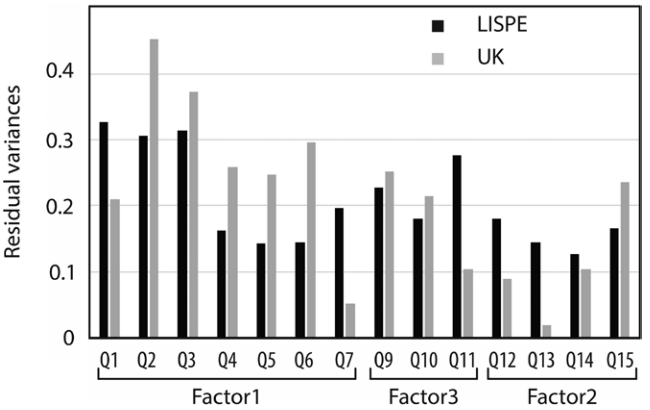

Fig. 1. Residual variances for each question (Q1–Q15) in exploratory factor analysis in Life-Space Mobility in Old Age (LISPE) and UK samples.

Factor Scores • For each participant, we calculated two types of factor scores: individually weighted and unitary-weighted factor scores. Summary statistics for each factor (products of observed scores and unitary/individual weights) as well as for the overall questionnaire score for unitary-weighted scores are shown in Table 3.

LISPE Sample • Mean scores ranged between 1.1 and 2.6 points for unitary-weighted and 0.5 and 1.9 points for individually weighted scores. Both types of scores were closely related across participants. Spearman correlation coefficients between the speech hearing factor, on the one hand, and the spatial hearing factor or the socio-emotional consequences factor, on the other hand, were 1.00, 0.94, and 0.98.

UK Sample • Mean scores ranged between 1.6 and 2.7 for the unitary-weighted and −0.4 and −0.2 for the individually weighted scores. Again, both types of scores were closely related across participants, with Spearman correlation coefficients between the speech hearing factor, on the one hand, and the spatial hearing factor or the socio-emotional consequences factor, on the other hand, 0.98, 0.90, and 0.98.

A comparison of group means between the LISPE and UK samples for unitary- and individually weighted scores showed no systematic differences between groups once *p* values were adjusted for false discovery rate. In the following, we will restrict reporting to unitary-weighted scores because they are more useful in cross-study comparisons.

TABLE 3. Descriptive statistics for each of the three factor scores

	Finnish LISPE Sample				UK Sample					
	n	M	SD	Range	*n*	M	SD	Range	p^a	Adjusted p^b
Unit score										
SpeH	535	2.6	2.2	0–8.7	50	2.3	1.8	0–7.9	0.572	0.806
SpatH	550	2.1	1.9	0–10.0	50	2.7	1.7	0–6.3	0.007	0.070
SocE	549	1.1	1.6	0–9.3	50	1.6	2.1	0–7.5	0.473	0.733
Overall	523	2.1	1.8	0–8.4	50	2.2	1.7	0–7.4	0.251	0.599
Ind score										
SpeH	523	0.55	3.7	−4.1 to 12	50	−0.4	3.4	−4.6 to 11	0.143	0.479
SpatH	523	0.5	2.1	−2.0 to 8.8	50	−0.03	3.0	−4.6 to 7.0	0.252	0.599
SocE	523	1.9	2.3	−0.2 to 12	50	−0.2	5.3	−5.4 to 17	0.025	0.155

[a]*Chi-square test p value between a model allowing different means for the samples and a model restricting the means to be equal. $p < 0.05$ indicates that the restricted model fits the data significantly worse.*

[b]*Adjusted by false discovery rate correction (q value).*

Unit score, unitary-weighted; M, mean; n, number of participants; SD, standard deviation; SocE, socio-emotional consequences; SpatH, spatial hearing; SpeH, speech hearing; ind score, individually weighted.

The correlations between unitary-loaded factors are given in Table 4. The results show two things. First, the correlations among factors are big enough to merit the use of an oblique rotation for the EFA. Second, the correlations of the spatial hearing factor with the speech hearing factor and the socio-emotional consequences factor were fairly similar across samples and indeed did not differ significantly (F3–F1, $p = 0.49$; F3–F2, $p = 0.73$; Steiger 1980). Conversely, the correlation between the socio-emotional consequences factor and the speech hearing factor differed significantly between the LISPE and UK samples ($p < 0.01$), suggesting more shared variance between the two factors in the LISPE than in the UK sample.

Internal Consistency

LISPE Sample • Cronbach's α was 0.96, 0.90, and 0.94 for Factors 1 to 3, respectively.

UK Sample • Cronbach's α was 0.95, 0.93, and 0.97, respectively.

The results show that the consistency scores are uniformly high for all three factors in both samples and well above the suggested cutoff value of 0.90 (Streiner 2003; Tavakol & Dennick 2011).

Test–Retest Reliability (COSMIN Points 3 and 4; UK Sample)

Test–retest reliability was assessed by means of $ICC_{3,1}$, SEM, and MD estimates. These parameters could only be calculated for the UK sample. The ICC, SEM, and MD values for the overall score and all three factor scores are presented in Table 5. ICC values ranged between 0.80 (socio-emotional consequences factor score) and 0.86 (overall score) and therefore satisfy test–retest reliability criteria (Cicchetti 1994). ICC, calculated using the formula assuming systematic error between the tests, yielded similar results. SEM values, reflecting typical expected error of the test score (SEM_s) and true score (SEM_{ts}), were lowest for the overall score and highest for the socio-emotional consequences score. Finally, the minimal difference required for a change in score to be meaningful varied between 2.35 points in the overall score and 3.29 points on the socio-emotional consequences subscale.

TABLE 4. GEOMIN factor correlations (SE) within the samples

	LISPE Sample		UK Sample	
	SpeH	SocE	SpeH	SocE
SpatH	0.61 (0.06)	0.66 (0.05)	0.54 (0.08)	0.63 (0.05)
SocE	0.72 (0.07)		0.46 (0.10)	

LISPE, Life-Space Mobility in Old Age; SocE, socio-emotional consequences; SpatH, spatial hearing; SpeH, speech hearing.

Hypothesis Testing: Testing the Relationship to Other Measurement Instruments and to Age (COSMIN Point 5; LISPE2 and UK Samples)

HERE, BEA, and Age

LISPE2 • The relationship of the three HERE factors as well as the overall HERE score (all unitary-weighted) to BEA and age was explored in a set of regression analyses (Table 6). Also considered was the relationship of Q1 to BEA and age on its own because Q1 ("How is your hearing?") is the direct self-report equivalent of the objective PTA measure (from which BEA is derived) and could be used on its own as a way to assess hearing sensitivity in samples for which PTA cannot be assessed. To be considered a significant predictor, the confidence interval of the predictor must exclude zero. The results show that BEA predicted a significant proportion of individual variability across all factors. It also predicted significant variance for the overall score and Q1 on its own. Age, on the other hand, did not predict a significant proportion of variance for any of the factors, overall score, or Q1. When computing the BEA as average of the

TABLE 5. HERE questionnaire test–retest reliability measures based on unitary-loaded mean factor scores (UK data)

	Test 1		Test 2					
	M	SD	M	SD	ICC	SEM_s	SEM_{ts}	MD
SpeH	2.3	1.8	2.4	1.9	0.84	1.02	0.94	2.59
SpatH	2.6	1.7	2.4	1.8	0.82	1.03	0.93	2.58
SocE	1.5	2.1	1.7	2.1	0.80	1.32	1.19	3.29
Overall	2.1	1.7	2.2	1.8	0.86	0.92	0.85	2.35

ICC, intraclass correlation coefficient; M, mean; MD, minimal difference; SD, standard deviation; SEM_s, standard error of measurement of the test score; SEM_{ts}, standard error of measurement of the true score; SocE, socio-emotional consequences; SpatH, spatial hearing; SpeH, speech hearing.

TABLE 6. Regression models estimating the predictive effect of (i) better ear pure-tone average over frequencies 0.25–8 kHz (BEA) and (ii) age for unitary-weighted HERE scores per factor and total unitary-weighted scores in the LISPE2 and UK sample

	LISPE2 Sample				UK Sample					
	B	95% CI	*β*	R^2	*B*	95% CI	*β*	R^2	χ^2 p[a]	Adjusted p[b]
BEA										
SpeH	0.12	0.08–0.15	0.51	0.26	0.11	0.07–0.15	0.57	0.33	0.808	0.917
SpatH	0.09	0.04–0.13	0.36	0.13	0.10	0.01–0.18	0.31	0.10	0.376	0.648
SocE	0.06	0.02–0.09	0.30	0.09	0.03	−0.01 to 0.07	0.18	0.03	0.812	0.917
Overall	0.08	0.05–0.11	0.45	0.20	0.08	0.04–0.12	0.46	0.21	0.928	0.928
Q1	0.13	0.08–0.17	0.51	0.26	0.12	0.07–0.16	0.56	0.31	0.695	0.898
Age										
SpeH	0.04	−0.01 to 0.09	0.07	0.005	0.07	−0.01 to 0.15	0.23	0.053	0.566	0.806
SpatH	0.02	−0.04 to 0.08	0.04	0.001	0.04	−0.14 to 0.22	0.09	0.007	0.346	0.631
SocE	0.02	−0.03 to 0.07	0.04	0.002	−0.02	−0.09 to 0.05	−0.07	0.004	0.844	0.917
Overall	0.03	−0.01 to 0.07	0.06	0.004	0.04	−0.04 to 0.11	0.13	0.017	0.886	0.917
Q1	0.05	−0.01 to 0.11	0.07	0.005	0.04	−0.05 to 0.12	0.12	0.014	0.825	0.917

[a]*p Value from the likelihood ratio test for comparison between model allowing unequal regression coefficients between the samples and model that fixes regression coefficients to be equal between the samples. $p < 0.05$ indicates that the fixed model fits the data significantly worse than the first model.*

[b]*Adjusted by false discovery rate correction (q value).*

B, unstandardized regression coefficient; β, standardized regression coefficient; BEA, better-ear pure-tone average; LISPE, Life-Space Mobility in Old Age; overall, score overall items; Q1, Question 1 in HERE questionnaire "How is your hearing?"; R^2, amount of variance accounted for by factor; SocE, socio-emotional consequences factor; SpatH, spatial hearing factor; SpeH, speech hearing factor.

most speech-relevant frequencies only (0.5 to 4 kHz), the results were very similar.

UK Sample • The results were very similar with one exception; BEA was not a significant predictor for the socio-emotional consequences factor. All other result patterns concerning BEA and age were as in the LISPE2 sample. The regression values between the two samples were remarkably similar. Therefore, it is not surprising that the chi-square tests showed no differences between the two samples.

HERE and SSQ_{speech} • In the UK sample only, we assessed construct validity by correlating the unitary-weighted scores of the speech hearing factor of the HERE questionnaire (7 questions) with the unitary-weighted scores of a comparable speech scale of another questionnaire, in this case the speech scale of the SSQ (14 questions). Note that the answer scales are reversed in the two questionnaires, with 0 indicating no difficulty in the HERE questionnaire and the highest degree of difficulty in the SSQ. We are therefore expecting the correlation to be negative. The Pearson product–moment correlation coefficient between the two scales was −0.75 ($p < 0.001$), indicating high covariance between responses to the speech-related questions in both questionnaires.

HERE and Speech Perception • Also in the UK sample, we assessed the Pearson product–moment correlation coefficient between the speech hearing factor of the HERE questionnaire and the average intelligibility score of a set of 112 sentences presented in background noise. Note again that a negative correlation between the two scores was to be expected as higher scores indicate more difficulty in the questionnaire but better intelligibility in the speech perception task. This correlation was −0.50 ($p < 0.001$). In contrast, the correlation between the SSQ_{speech} and the same speech test was only $r = 0.29$ ($p < 0.05$), significantly lower according to a direct comparison ($p < 0.05$; Steiger 1980). Q1 on its own correlated with the sentence intelligibility score at −0.43 ($p < 0.05$).

DISCUSSION

The psychometric quality of hearing-related questionnaires is often insufficiently assessed. This puts the usability of a questionnaire into serious doubt. COSMIN (Mokkink et al. 2010), a four-round Delphi study, provides consensus-based standards that health status measurement instruments need to fulfill to be deemed useful. In this study, we aimed to demonstrate how the psychometric quality of a questionnaire can be evaluated by following the COSMIN criteria. We do this by estimating the psychometric properties of the newly developed HERE questionnaire. By doing so, we hope to provide a short and well-validated measurement instrument to those who wish to assess hearing function and its socio-emotional consequences in social situations and environments.

In the following section, we will discuss each psychometric aspect suggested by COSMIN as essential for psychometric validation. Because the comparison of two demographically and culturally very different samples is a central aspect of this study, intersample differences (COSMIN Point 6) are discussed in relation to every other aspects of psychometric validity.

Structural Validity (COSMIN Points 1 and 2)

Questionnaire Scores • Questionnaire scores between the two samples were broadly comparable. The only significant difference on the level of individual questions was greater reported difficulty estimating direction and location of sound sources in the UK sample (Q11). The absence of a difference between samples in Q1 is somewhat unexpected given the demographic differences between samples (the LISPE sample was on average 13 years older than the UK sample and had objectively poorer hearing thresholds as reflected by higher BEA scores) and given that objective (BEA) and subjective (Q1) assessment of hearing sensitivity was correlated in the two samples (0.51 in the LISPE and 0.56 in the UK sample). It would have been reasonable to expect LISPE participants to report more subjective difficulty with hearing. In a similar vein, despite objectively poorer hearing for the Finnish participants, they did not perceive themselves as any more impaired on various aspects of speech perception and, possibly as a consequence, did not perceive themselves as suffering disproportionately from socio-emotional consequences of their objective hearing loss.

A number of interpretations are possible for this result. One concerns the changing relationship of age with stigma. Erler and Garstecki (2002) showed that stigma related to hearing loss was perceived as higher among younger (35 to 45 and 55 to 65) than older women (75 to 85). Possibly, then, the comparatively fewer negative socio-emotional consequences of their objectively poorer hearing were due to the reduced stigmatization of their hearing loss in the older adults in the LISPE group. Another possible interpretation is a response shift adjustment for the older group as suggested by Treadwell and Lenert's (1999) Prospect Theory. This theory posits two tenets: first, a worsening health status leads to a decrease in perceived health, which is first steep and then levels off as the decline continues; second, their current health status provides an individual with a reference point for their standard of health, and people with poorer health (in this case hearing loss) tend to have a lower standard for good health than people with better health (i.e., those with minimal hearing loss). This change in standard, called response shift (Howard 1980), occurs when a person has become accustomed to their permanent health change (Treadwell & Lenert 1999). A response shift typically has two consequences: first, it underestimates the measured phenomenon; second, it adjusts the gap between the perceived optimal and present state based on the permanently reduced health state. In the case of hearing, it would mean that the same amount of hearing loss leads to greater perceived difficulties in people who have not yet experienced the response shift due to a more recent onset of their hearing loss. Often, these people will also be younger. Empirical evidence for this interpretation is provided by Gordon-Salant et al. (1994) who found that younger (<40 years) persons with hearing loss reported more hearing disability (HHIE) than older persons with similar audiograms. Hence, by comparison, the psychosocial effects of hearing difficulties are rated as less severe in the group that has experienced the response shift (older adults). In the current study, this would mean that the older adults in the LISPE sample have experienced the response shift and as a consequence rate their psychosocial effects no differently to the less impaired group.

The only significant difference in scores between samples in the study were the higher scores for perceived difficulty of spatial hearing in the UK compared to LISPE sample. This was despite UK participants having objectively better hearing both in terms of better overall sensitivity and lower interaural differences. Moreover, the subjective difficulties in the UK sample were entirely unrelated to hearing sensitivity, both overall (BEA) and in terms of interaural differences. In contrast, the LISPE2 sample showed significant correlations between the response to Q9 to Q11 and interaural differences (Spearman $r = 0.20$ to 0.24). The correlations indicated that greater interaural differences were associated with greater difficulty in spatial hearing. Hence, as LISPE participants had greater interaural differences, they should have rated their spatial hearing as poorer than UK participants. We interpret these results again in the context of Prospect Theory and assume that the older LISPE2 participants might have experienced a shift in reference point, and this shift might have been more pronounced for listeners with worse hearing and greater interaural differences.

In terms of why UK participants had slightly higher scores for Q9 to Q11, it is possible that the UK participants, who were also younger, enjoyed better mobility and more physical activity outdoors, which in turn enabled them to enter more situations that required good spatial auditory functioning and made difficulties more noticeable. This potential difference between samples is supported by the location in which the testing was accomplished: UK participants were required to travel to the laboratory to take part in the study, while the LISPE participants were tested in their homes, and thus the sample included participants with limitations in outdoor mobility. Potentially, the inclusion of questions designed to assess hearing in situations which required outdoor mobility may have highlighted differences between samples.

Factor Modeling • In the present study, we showed the internal structure of the HERE to be similar for cohorts from two different countries, Finland and the United Kingdom. In both samples, Q1 to Q7 loaded highly on the speech hearing factor, Q9 to Q11 on the spatial hearing factor, and Q12 to Q15 on the socio-emotional consequences factor. In contrast to the Finnish sample, the smaller UK sample showed substantial secondary loadings on another factor for all but one question (Q2) on the Speech Hearing factor (Factor1). For Q1, the secondary loading was with spatial hearing, which may indicate how important spatial hearing was to the UK sample when evaluating their overall hearing ability. For Q3 to Q7, the secondary loadings were with the socio-emotional consequence factor, indicating how perceived difficulties in speech perception were closely linked to socio-emotional consequences.

While the overall fit of the factor structure as assessed by communality values was comparable for the two samples, the fit of individual factors varied. Specifically, the speech hearing factor seemed to have less residual variability, that is, a better fit in the LISPE than in the UK sample. Conversely, the spatial hearing and socio-emotional consequences factors appeared to fit equally well in both samples as indicated by similar factor weights, communalities, and residual variances. Generally, however, it is remarkable how similar the relationship between surface questions and latent factors was in the two samples despite differences in hearing sensitivity, culture, and activity. This similarity of factor structure is further substantiated by a previous study by Polku et al. (2018), who confirmed the same factor structure for a sample of Finnish participants that consisted of both hearing aid users and non-hearing aid users and that used best-hearing score (with or without a hearing aid) as the hearing variable input for an EFA. All of these results suggest that the questionnaire items measure similar underlying concepts in these two populations. Because subtle differences in internal factor structure might exist due to the particular makeup of a sample or subtle changes introduced by the translation between languages, we used unitary-weighted scores. As unitary-weighted scores do not mitigate against measurement error, they provide no guarantee that the similarities of factor structures between samples relate to hearing dimensions; alternatively, they could relate to similarities based on other sources.

Internal Consistency • The internal consistency of the three factors was uniformly high across both samples, indicating high inter-item covariances, which makes it probable that the questions assigned to a factor assessed the same underlying concept.

Test–Retest Reliability and Measurement Error (COSMIN Points 3 and 4)

ICC ranged between 0.80 and 0.86 for all three factors and the overall score, which according to Cicchetti (1994) indicates

excellent test–retest reliability. These high values are particularly remarkable given the longtime interval between test and retest (mean of 184 days). Such a longtime interval reduces the likelihood of responses being recalled from the first test rather than being reassessed at the time of retesting, a concern that may be present when the time interval is very short. We also found no evidence of a systematic change in scores in our rather long test interval, which raises the possibility that this questionnaire could be suitable for intervention monitoring as its scoring is stable over time. However, whether the HERE questionnaire is sensitive to change over time remains to be investigated in future studies. Our results are comparable to the test–retest reliability of other questionnaires. For instance, a test–retest ICC of 0.85 in a one-way random analysis was found by Tomioka et al. for the HHIE-S in a Japanese cohort of older adults (Tomioka et al. 2013). Weinstein et al. (2015) found a test–retest ICC of 0.99 in an Arabic version of the HHIE-S; however, they did not report which ICC formula they used. The exceptionally high ICC may be explained by the very short time interval, 1 hour, between the tests, where it is possible that participants recalled their answers to the first test rather than freshly assessing their hearing ability in the retest setting. Finally, Singh and Pichora-Fuller (2010) found ICCs between 0.65 and 0.83 for the SSQ questionnaire, with the highest correlations restricted to situations where an interview method was used at both test times. Other studies have either used Pearson product–moment correlation coefficients instead of ICCs (Weinstein et al. 1986; Lichtenstein et al. 1988; Newman et al. 1991) or have not reported which correlation coefficient they used (Noble et al. 2012; correlation coefficient 0.54); this makes it difficult to compare them to the present study, as it has been shown that the formula used may have considerable effect on the magnitude of the ICC (Moulin et al. 2015).

We assessed the standard error of measurement as a way to estimate the minimum difference in score needed for a change to be meaningful for an individual. This minimum difference varied between 2.4 and 3.3 points between scales. Given how stable the HERE score are across time when measured without intervention, the knowledge of the minimum difference is a potentially useful piece of information as it provides a target for score change over time due to intervention.

The Relationship to Other Measurement Instruments (COSMIN Point 5)

We tested the predictive value of age and hearing sensitivity for HERE scores. A set of regression analyses showed that, despite differences in age and hearing level (BEA) between the groups, the predictive relationship between these two variables and either factor or total HERE scores was very similar for the two samples, suggesting that the questionnaire is not responsive to age-related differences and shows the same relationship to hearing sensitivity regardless of differences in hearing level. In the case of age, no predictive relationship was found for either sample. In the case of hearing level, HERE scores were significantly associated with behaviorally measured hearing sensitivity (BEA) in both samples and explained about 20% of variability overall. This compares well with the finding by Chang et al. (2009) that 27% of variance in the HHIE-S was explained by BEA (0.5 to 4 kHz) but is substantially lower than the result obtained in a Japanese study (Tomioka et al. 2013), where BEA (0.5 to 4kHz) explained 48% of the variance in HHIE-S.

We also investigated the relationship between BEA and Q1 on its own because the question represents the clearest equivalence between subjectively assessed hearing function and BEA. Surprisingly, the correlation between BEA and perceived hearing was almost identical regardless of whether the perceived hearing score was based on the one question "How is your hearing?" or a range of questions aimed at assessing hearing across a variety of situations. This result suggests that if a study is looking to include self-report questions as a proxy measure of hearing, Q1 on its own would be as appropriate as including all seven questions of the speech hearing factor.

We also assessed the relationship of the HERE to other measures that assess speech perception, specifically to one self-report (SSQ_{Speech} subscale) and one behavioral (intelligibility of sentences in noise) measurement. We found a high correlation with the SSQ_{Speech} subscale, suggesting that both self-report measures assess a similar construct. Regarding the behavioral measure, we found a moderately high correlation ($r = -0.50$) between the HERE's speech hearing factor and the speech intelligibility measure. Note that this correlation was significantly higher than the correlation between the SSQ_{Speech} subscale and the intelligibility measure, making the HERE potentially more useful as a substitute for a behavioral speech-in-noise perception test. This result also suggests that the variance which connects the HERE's speech hearing scores and SSQ_{Speech} scores is not driven by a shared variance with the speech-in-noise measure.

Q1 on its own showed a somewhat lower correlation to speech perception than the combined speech hearing score of the HERE. Researchers looking for a proxy measure of speech perception might have to choose between a more accurate representation of behavioral speech perception using seven questions or a faster but less accurate assessment using Q1 only (for similar results see also Nondahl et al. 1998; Salonen et al. 2011).

Usability

Its brevity, good test–retest reliability, and high correlation with a behavioral speech-in-noise perception measure make the HERE questionnaire useable for a range of applications. We successfully used the HERE in a population-based cohort study where questionnaires were administered via postal mailing without any need for supervision. The questionnaires, when returned, had been filled in correctly, despite some missing responses. Hence, the HERE questionnaire can be used even in large surveys.

The use of a numeric response scale, in contrast to a three-response categories scale which leads most people to choose "sometimes" or "some difficulty," provided sufficient variability in the responses to conduct advanced statistical analyses. Based on previous research (Akeroyd et al. 2014) which showed that a continuous scale would not have added value by producing continuous data, because respondents tend to choose integers on a continuous scale, we chose a numeric rating scale. The distribution of the responses showed left censoring which needed to be taken into account in the statistical analyses. This is also likely to be the case in other hearing questionnaires when applied in nonpatient populations (although most studies do not describe the distribution). The HERE questionnaire makes it possible to compare hearing with and without a hearing aid in hearing aid users, and it is possible to analyze "best hearing" in cohort studies (i.e., hearing with or without a hearing aid, whichever is

better). We believe that this reflects hearing in everyday situations, as people are likely to use hearing aids primarily in situations in which they experience benefit from the aid. Through its explicit link between behavioral and self-report measures of speech perception, this study demonstrates how experimental studies aiming to understand mechanisms underlying speech perception can be related to large-scale epidemiological studies that investigate speech perception in the real world.

Because in the context of the current study none of the existing hearing questionnaires satisfied all five inclusion criteria, we designed a new questionnaire and validated it using the COSMIN criteria. We do not wish to claim that the HERE is more useful than any of the other popularly used questionnaires. Rather, we wish to make the point that regardless of whether a questionnaire is old or new, it needs to be validated concerning all COSMIN criteria. We also wish to point out that there are a variety of statistical tools available to achieve this. We use a particular selection and give our reasoning for our choice. We also provide some practical information concerning these tools in case others want to use them to validate a questionnaire of their choice.

Study Limitations

The LISPE sample probably overrepresented the well-functioning subsection of the population. The UK sample was small for the EFA. This makes it hard to know whether the differences found in the internal questionnaire structure between the samples were due to genuine differences between samples or due to a potentially greater sampling error in the smaller sample. The cross-cultural comparison was not optimal as it utilized two samples that were not totally comparable in terms of age and hearing sensitivity. Hearing sensitivity was measured in a laboratory setting in the UK sample, whereas in LISPE it was measured at the participants' homes, which may cause some differences in observed hearing levels. Speech intelligibility and SSQ scores were only available for the UK sample, and thus any analyses including one of these measures was limited to the UK sample. Although we assume that results would be similar in the Finnish sample, this remains to be demonstrated.

ACKNOWLEDGMENT

We thank our HEARATTN (Hearing, remembering and living well: Paying attention to challenges that older adults face in noisy environments) consortium partners for their contributions in the development of HERE questionnaire: Bruce Schneider, PhD, University of Toronto, Canada; Jean-Pierre Gagné, PhD, Researcher, Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Université de Montréal, Montréal, Québec, Canada; Daniel A. Levy, PhD, and Boaz Ben-David, PhD, The Interdisciplinary Center Herzliya, Israel. The authors also thank Sarah Knight for help with data collection of the UK sample and Oliver Zobay for valuable guidance in the statistical analyses.

A. V., T. M. M., and H. P. designed the HERE questionnaire and collected the Finnish data. A. H. designed all experiments and collected all data for the UK project. T. M. M., A. H., and T. T. analyzed and interpreted the data. A. H. and T. M. M. wrote the manuscript with critical input from A. V., T. T., and H. P.

A. H. and T. M. M. contributed equally to this work.

This research was funded by grant BB/K021508/1 from the Biotechnology and Biological Sciences Research Council (to A. H.), grant U135097128 from the Medical Research Council (to A. H.), grant 263729 from the Academy of Finland (to A. V.), and from the Juho Vainio Foundation (to H. P.).

Part of the results were presented at Gerontologia 2017 in Turku, Finland.

The authors have no conflicts of interest to disclose.

Address for correspondence: Antje Heinrich, Manchester Centre for Audiology and Deafness (ManCAD), School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, M13 9PL, United Kingdom. E-mail: antje.heinrich@manchester.ac.uk

Received April 27, 2017; accepted April 23, 2018.

REFERENCES

Akeroyd, M. A., Guy, F. H., Harrison, D. L., et al. (2014). A factor analysis of the SSQ (Speech, Spatial, and Qualities of Hearing Scale). *Int J Audiol*, *53*, 101–114.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B*, *57*, 289–300.

Besser, J., Zekveld, A. A., Kramer, S. E., et al. (2012). New measures of masked text recognition in relation to speech-in-noise perception and their associations with age and cognitive abilities. *J Speech Lang Hear Res*, *55*, 194–209.

Bolarinwa, O. A. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Niger Postgrad Med J*, *22*, 195–201.

British Society of Audiology. (2011). Recommended Procedure: Pure-Tone Air-Conduction and Bone-Conduction Threshold Audiometry With and Without Masking. Retrieved from http://www.thebsa.org.uk/wp-content/uploads/2011/04/Pure-Tone-Audiometry.pdf. on June 23, 2014.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*, *56*, 81–105.

Chang, H. P., Ho, C. Y., Chou, P. (2009). The factors associated with a self-perceived hearing handicap in elderly people with hearing impairment–results from a community-based study. *Ear Hear*, *30*, 576–583.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*, *6*, 284–290.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract Assess Res Eval*, *10*, 1–9.

Cox, R. M., & Alexander, G. C. (1995). The abbreviated profile of hearing aid benefit. *Ear Hear*, *16*, 176–186.

Cruice, M., Worrall, L., Hickson, L. (2006). Quantifying aphasic people's social lives in the context of non-aphasic peers. *Aphasiology*, *20*, 1210–1225.

Divenyi, P. L., & Haupt, K. M. (1997). Audiological correlates of speech understanding deficits in elderly listeners with mild-to-moderate hearing loss. I. Age and lateral asymmetry effects. *Ear Hear*, *18*, 42–61.

Duquesnoy, A. J. (1983). The intelligibility of sentences in quiet and in noise in aged listeners. *J Acoust Soc Am*, *74*, 1136–1144.

Era, P., Jokela, J., Qvarnberg, Y., et al. (1986). Pure-tone thresholds, speech understanding, and their correlates in samples of men of different ages. *Audiology*, *25*, 338–352.

Erler, S. F., & Garstecki, D. C. (2002). Hearing loss- and hearing aid-related stigma: Perceptions of women with age-normal hearing. *Am J Audiol*, *11*, 83–91.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess*, *7*, 286–299.

Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *Int J Audiol*, *43*, 85–99.

Gordon-Salant, S., Lantz, J., Fitzgibbons, P. (1994). Age effects on measures of hearing disability. *Ear Hear*, *15*, 262–265.

Hall, D. A., Zaragoza Domingo, S., Hamdache, L. Z., et al.; International Collegium of Rehabilitative Audiology and TINnitus Research NETwork. (2018). A good practice guide for translating and adapting hearing-related questionnaires for different languages and cultures. *Int J Audiol*, *57*, 161–175.

Heinrich, A., Gagné, J.-P., Viljanen, A., et al. (2016a). Effective communication as a fundamental aspect of active aging and well-being: Paying attention to the challenges older adults face in noisy environments. *Social Inquiry Into Well-Being*, *2*, 51–69.

Heinrich, A., Henshaw, H., Ferguson, M. A. (2015). The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests. *Front Psychol*, *6*, 782.

Heinrich, A., Henshaw, H., Ferguson, M. A. (2016b). Only behavioral but not self-report measures of speech perception correlate with cognitive abilities. *Front Psychol*, *7*, 576.

Helfer, K. S., & Wilber, L. A. (1990). Hearing loss, aging, and speech perception in reverberation and noise. *J Speech Hear Res*, *33*, 149–155.
Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Med*, *30*, 1–15.
Horn, J. I. (1965). A rationale and a test for the number of factors in factor analysis. *Psychometika*, *30*, 179–185.
Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Eval Rev*, *4*, 93–106.
Humes, L. E., & Christopherson, L. (1991). Speech identification difficulties of hearing-impaired elderly persons: the contributions of auditory processing deficits. *J Speech Hear Res*, *34*, 686–693.
Humes, L. E., & Roberts, L. (1990). Speech-recognition difficulties of the hearing-impaired elderly: the contributions of audibility. *J Speech Hear Res*, *33*, 726–735.
Humes, L. E., Watson, B. U., Christensen, L. A., et al. (1994). Factors associated with individual differences in clinical measures of speech recognition among the elderly. *J Speech Hear Res*, *37*, 465–474.
IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp; 2013.
Jerger, J., & Chmiel, R. (1997). Factor analytic structure of auditory impairment in elderly persons. *J Am Acad Audiol*, *8*, 269–276.
Jerger, J., Jerger, S., Pirozzolo, F. (1991). Correlational analysis of speech audiometric scores, hearing loss, age, and cognitive abilities in the elderly. *Ear Hear*, *12*, 103–109.
Kamakura, W. A., & Wedel, M. (2001). Exploratory tobit factor analysis for multivariate censored data. *Multivar Behav Res*, *36*, 53–82.
Kramer, S. E., Kapteyn, T. S., Kuik, D. J., et al. (2002). The association of hearing impairment and chronic diseases with psychosocial health status in older age. *J Aging Health*, *14*, 122–137.
Lichtenstein, M. J., Bess, F. H., Logan, S. A. (1988). Validation of screening tools for identifying hearing-impaired elderly in primary care. *JAMA*, *259*, 2875–2878.
Mikkola, T. M., Polku, H., Sainio, P., et al. (2016). Hearing loss and use of health services: A population-based cross-sectional study among Finnish older adults. *BMC Geriatr*, *16*, 182.
Mokkink, L. B., Terwee, C. B., Patrick, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*, *19*, 539–549.
Moulin, A., Pauzie, A., Richard, C. (2015). Validation of a French translation of the Speech, Spatial, and Qualities of Hearing Scale (SSQ) and comparison with other language versions. *Int J Audiol*, *54*, 889–898.
Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus User's Guide (6th ed.). Los Angeles, CA: Muthén & Muthén.
Newman, C. W., Weinstein, B. E., Jacobson, G. P., et al. (1991). Test-retest reliability of the hearing handicap inventory for adults. *Ear Hear*, *12*, 355–357.
Noble, W., Naylor, G., Bhullar, N., et al. (2012). Self-assessed hearing abilities in middle- and older-age adults: a stratified sampling approach. *Int J Audiol*, *51*, 174–180.
Nondahl, D. M., Cruickshanks, K. J., Wiley, T. L., et al. (1998). Accuracy of self-reported hearing loss. *Audiology*, *37*, 295–301.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. (3rd ed.). New York: McGraw-Hill.
Polku, H., Mikkola, T. M., Rantakokko, M., et al. (2018). Hearing and quality of life among community-dwelling older adults. *J Gerontol B Psychol Sci Soc Sci*, *73*, 543–552.
Rantanen, T., Portegijs, E., Viljanen, A., et al. (2012). Individual and environmental factors underlying life space of older people—Study protocol and design of a cohort study on life-space mobility in old age (LISPE). *BMC Public Health*, *12*, 1018.
Salonen, J., Johansson, R., Karjalainen, S., et al. (2011). Relationship between self-reported hearing and measured hearing impairment in an elderly population in Finland. *Int J Audiol*, *50*, 297–302.
Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans, D. S. G. Pollock, A. Satorra (Eds.), *Innovations in Multivariate Statistical Analysis: A Festschrift for Heinz Neudecker* (pp. 233–237). Dordrecht, The Netherlands: Kluwer Academic Publishers.
Singh, G., & Pichora-Fuller, M. (2010). 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*, *49*, 733–740.
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–253.
Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *J Pers Assess*, *80*, 99–103.
Studebaker, G. A. (1985). A "rationalized" arcsine transform. *J Speech Hear Res*, *28*, 455–462.
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *Int J Med Educ*, *2*, 53–55.
Tomioka, K., Ikeda, H., Hanaie, K., et al. (2013). The Hearing Handicap Inventory for Elderly-Screening (HHIE-S) versus a single question: Reliability, validity, and relations with quality of life measures in the elderly community, Japan. *Qual Life Res*, *22*, 1151–1159.
Treadwell, J. R., & Lenert, L. A. (1999). Health values and prospect theory. *Med Decis Making*, *19*, 344–352.
van Rooij, J. C., & Plomp, R. (1990). Auditive and cognitive factors in speech perception by elderly listeners. II: Multivariate analyses. *J Acoust Soc Am*, *88*, 2611–2624.
van Rooij, J. C., & Plomp, R. (1992). Auditive and cognitive factors in speech perception by elderly listeners. III. Additional data and final discussion. *J Acoust Soc Am*, *91*, 1028–1033.
van Rooij, J. C., Plomp, R., Orlebeke, J. F. (1989). Auditive and cognitive factors in speech perception by elderly listeners. I: Development of test battery. *J Acoust Soc Am*, *86*, 1294–1309.
Ventry, I. M., & Weinstein, B. E. (1982). The hearing handicap inventory for the elderly: A new tool. *Ear Hear*, *3*, 128–134.
Viljanen, A., Törmäkangas, T., Vestergaard, S., et al. (2014). Dual sensory loss and social participation in older Europeans. *Eur J Ageing*, *11*, 155–167.
Weinstein, B. E., Rasheedy, D., Taha, H. M., et al. (2015). Cross-cultural adaptation of an Arabic version of the 10-item hearing handicap inventory. *Int J Audiol*, *54*, 341–346.
Weinstein, B. E., Spitzer, J. B., Ventry, I. M. (1986). Test-retest reliability of the Hearing Handicap Inventory for the Elderly. *Ear Hear*, *7*, 295–299.
Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*, *19*, 231–240.
Whitmer, W. M., Wright-Whyte, K. F., Holman, J. A., et al. (2015). Hearing aid validation. In B. C. J. Moore (Ed.), *Springer Handbook of Auditory Research: Hearing Aids* (pp. 291-321). New York: Springer-Verlag.

REFERENCE NOTE

1. Muthén, B. O., du Toit, S. H. C., Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished technical report. Available at https://www.statmodel.com/download/Article_075.pdf.