AMERICAN ACADEMY OF
OTOLARYNGOLOGY–
HEAD AND NECK SURGERY
FOUNDATION

Otolaryngology–
Head and Neck Surgery
2018, Vol. 158(6) 1093–1100
© American Academy of
Otolaryngology–Head and Neck
Surgery Foundation 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0194599818760528
http://otojournal.org
SAGE

Improving Measurement Efficiency of the Inner EAR Scale with Item Response Theory

Annika Jessen, EdM[1], Andrew D. Ho, PhD[1],
C. Eduardo Corrales, MD, PhD[2], Bevan Yueh, MD, MPH[3],
and Jennifer J. Shin, MD, SM[2]

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

Abstract

Objectives. (1) To assess the 11-item Inner Effectiveness of Auditory Rehabilitation (Inner EAR) instrument with item response theory (IRT). (2) To determine whether the underlying latent ability could also be accurately represented by a subset of the items for use in high-volume clinical scenarios. (3) To determine whether the Inner EAR instrument correlates with pure tone thresholds and word recognition scores.

Design. IRT evaluation of prospective cohort data.

Setting. Tertiary care academic ambulatory otolaryngology clinic.

Subjects and Methods. Modern psychometric methods, including factor analysis and IRT, were used to assess unidimensionality and item properties. Regression methods were used to assess prediction of word recognition and pure tone audiometry scores.

Results. The Inner EAR scale is unidimensional, and items varied in their location and information. Information parameter estimates ranged from 1.63 to 4.52, with higher values indicating more useful items. The IRT model provided a basis for identifying 2 sets of items with relatively lower information parameters. Item information functions demonstrated which items added insubstantial value over and above other items and were removed in stages, creating a 8- and 3-item Inner EAR scale for more efficient assessment. The 8-item version accurately reflected the underlying construct. All versions correlated moderately with word recognition scores and pure tone averages.

Conclusion. The 11-, 8-, and 3-item versions of the Inner EAR scale have strong psychometric properties, and there is correlational validity evidence for the observed scores. Modern psychometric methods can help streamline care delivery by maximizing relevant information per item administered.

Keywords

inner ear, hearing, hearing loss, health status, quality of life, validated instrument, psychometrics, item response theory, factor analysis, clinical care, audiometry, word recognition scores, pure tone averages

Received October 8, 2017; revised January 8, 2018; accepted January 31, 2018.

Whether motivated by interest in longitudinal outcomes research or by participation in the Centers for Medicare and Medicaid Services' Quality Payment Program, physicians have been increasingly drawn to collecting patient-reported assessments of health status.[1-3] These patient-centered metrics are ideally tracked with validated instruments that have demonstrated interrater reliability, internal consistency, discriminant validity, and responsiveness to change.[4,5] Such psychometrically validated instruments have traditionally been developed for broad-scale survey research, and their use in daily clinical practice has been expanding.[4] Accordingly, a growing cadre of providers and organizations seek to incorporate patient-reported assessments at regular clinical intervals, prompting increased interest in identifying the optimal balance of data acquisition and concision in administration.[3] Thus, methods for optimizing the efficiency of key instruments have direct clinical applicability. These methods include item response theory (IRT) and factor-analytic approaches to improve outcome measurements and the efficiency of large-scale data collection.[6,7]

IRT provides information about each instrument item as it relates to a latent ability or proficiency. IRT was popularized by Lord[8] and has been used to develop many widely used educational tests, such as the SAT and National Assessment of Educational Progress.[9] In these applications,

[1]Harvard Graduate School of Education, Cambridge, Massachusetts, USA
[2]Department of Otolaryngology, Harvard Medical School, Boston, Massachusetts, USA
[3]Department of Otolaryngology, University of Minnesota Medical School, Minneapolis, Minnesota, USA

Corresponding Author:
Jennifer J. Shin, MD, SM, Department of Otolaryngology, Harvard Medical School, 45 Francis Street, Boston, MA 02115, USA.
Email: jennifer_shin@meei.harvard.edu

Table 1. Item Means (SD) and Interitem Correlations (n = 334-353).

	Item										
	ear11	ear2a	ear2b	ear2c	ear2d	ear2e	ear3a	ear3b	ear4	ear5a	ear5b
Mean (0-100)	56.23	41.22	50.71	26.91	31.44	47.03	36.15	39.06	47.59	48.36	68.21
SD (0-100)	23.700	25.020	25.340	23.570	25.050	27.080	24.130	28.700	28.850	31.940	35.040
ear11: Ability to hear	1.00										
ear2: Ability to understand . . .											
Family and close friends (ear2a)	0.62	1.00									
Speech in a quiet room (ear2b)	0.49	0.74	1.00								
Conversation in a crowded restaurant (ear2c)	0.49	0.69	0.59	1.00							
Filter out unwanted noises (ear2d)	0.48	0.72	0.59	0.80	1.00						
Telephone conversations (ear2e)	0.41	0.63	0.61	0.51	0.60	1.00					
ear3: Ability to hear . . .											
In different listening situations (ear3a)	0.54	0.76	0.70	0.72	0.74	0.68	1.00				
Soft household sounds (ear3b)	0.47	0.65	0.68	0.56	0.61	0.72	0.70	1.00			
ear4: Mood based on your ability to hear	0.45	0.59	0.53	0.52	0.59	0.62	0.67	0.64	1.00		
ear5: Bother from . . .											
Need for repetition (ear5a)	0.33	0.52	0.39	0.52	0.52	0.42	0.54	0.48	0.58	1.00	
Restricted activities because of hearing (ear5b)	0.42	0.47	0.42	0.44	0.44	0.39	0.50	0.46	0.59	0.65	1.00

IRT characterizes a single dimension, such as math or reading proficiency (although IRT has the capacity to model multiple dimensions).[9] IRT improves on "classical test theory," whose item statistics (eg, percentage of questions about math answered correctly) depend on the population being tested (eg, astrophysicists or kindergarteners).[10] IRT instead describes question characteristics that are expected to remain constant, regardless of the respondent population or the questions that accompany it. These item parameters include information and location. *Information* describes how well an item can distinguish among patients, and *location* describes where on the ability scale an item provides this information. Together, they characterize the usefulness of each question in a validated instrument. These location and information parameters remain stable so long as the proposed IRT model fits the data.[11] This stability is beneficial, as it can provide a strong basis for streamlined testing with fewer, higher-yield items. IRT can also be used to generate "item maps," which depict item responses for given overall scores. Item maps may be helpful to clinicians who find overall scores too abstract; these maps show specific questions/answers as they relate to latent ability levels. These insights can then be used to support individualized clinical care.

Hearing loss is a frequently presenting condition, which affects 19% of adolescents and 60% of septuagenarians.[12,13] Its impact has made it a focus of Healthy People 2020 and multiple professional organizations.[14-16] Measuring this impact across large populations is thus important and may help determine which interventions improve patient care. To further our ability to optimally and efficiently track hearing-specific outcomes, we fit an IRT model to responses to the Inner Effectiveness of Auditory Rehabilitation (Inner EAR) instrument, a scale designed to assess hearing-related function and quality of life.[17] We also used factor analysis to assess the dimensionality of the scale. Our research questions were as follows:

- Does the Inner EAR scale measure a single factor (perceived hearing ability) as designed?
- What are the IRT location and information parameter estimates for each item? With these in mind, can a subset of items compose a shorter scale while preserving accuracy and validity of the score?
- Is there evidence of correlational validity between audiometric results and the original Inner EAR version or proposed shortened scale scores?

Methods

Data Collection

The protocol was approved by the Partners Institutional Review Board. Consecutive adult patients ($\geq$18 years) who presented with a self-selected chief complaint of hearing loss were offered the Inner EAR instrument. Completion occurred electronically after arrival to the participating hospital-based otolaryngology clinic between July 2014 and September 2016.[18]

Inner EAR Instrument

The Inner EAR scale assesses hearing-related function and quality of life; it was designed to help inform the decision for amplification, as well as to compare the impact of varying hearing-aid technologies on daily function. The scale contains 11 items, each of which was coded as delineated in **Table 1**.[17] The first of these items (ear11) provides a global overview: patients rate their overall ability to hear on an 11-point scale (0 = hate it, 5 = it's OK, 10 = love it). In the subsequent 10 items, patients rate their ability to understand (5 items: ear2a-ear2e), their ability to hear (2 items: ear3a, ear3b), their mood (1 item: ear4), and how bothered they are by their hearing (2 items: ear5a, ear5b). A 5-point Likert scale was used for the first 8 items (1 = poor, 5 = excellent) and a 4-point Likert scale for the final 2 items (1 = very bothered, 4 = not bothered). Collectively, this instrument aims to provide otolaryngologists and primary care providers with an indication of hearing-related function and quality of life.[17]

Based on the original validation study, calculation of the composite score for the 11-item Inner EAR scale excludes the first item (the global scale for overall ability to hear).[17] The remaining individual items are scored on a 0-100 scale. Here, we evaluated all 11 items to estimate the information that they provide empirically. In the subsequent analysis, we used average scores (on a 100-point scale) for classical test theory analyses and estimated scale scores for IRT analyses.

Audiometry

We assessed the correlational validity of Inner EAR results against formal audiometry; all patients who were newly presenting underwent hearing testing. Audiometric testing was performed with standard techniques, evaluating air and bone conduction thresholds at 250, 500, 1000, 2000, 4000, and 8000 Hz. Pure tone averages (PTAs) were computed with the thresholds at 500, 1000, and 2000 Hz. Word recognition scores (WRSs) were assessed as the percentage of correctly identified words, with recorded W-22 25-word lists (Central Institute for the Deaf, St Louis, Missouri). An audiometer (GSI 61; Grason-Stadler, Eden Prairie, Minnesota) with TDH headphones (Telephonics, Farmingdale, New York) or EAR insert phones (Etymotic Research, Elk Grove Village, Illinois) were utilized.

Data Analysis

Analysis 1 evaluated dimensionality, meaning whether the data indicated 1 underlying ability (eg, hearing ability) or multiple underlying abilities (eg, hearing ability and possibly nonhearing-related quality of life). Cronbach's alpha was calculated to assess internal consistency/reliability. The fit of a unidimensional model to the data was evaluated, and the percentage of total variance in scores that was attributable to the first principal component was determined. Together, these analyses can build a case for unidimensionality and reliability—a single construct, measured precisely.

In analysis 2, an IRT model suited for Likert-scaled items known as the *graded response model* (see Appendix 1 in the online version of the article) was fit to the data to estimate location and information parameters for each item. Items that provided less information were then considered for removal, resulting in 2 proposed shortened versions of the Inner EAR scale. An 8-item version included items related to perceived ability to filter noises, ability to hear in different situations, and mood based on hearing ability. A 3-item version included items related to understanding of family and friends, perceived ability to filter noises, and listening in varying environments. IRT also enabled item maps that can aid interpretation of scores. These maps illustrate how likely patients are to respond to each item, given their summary scores,[19] so clinicians can efficiently understand patient needs.

Analysis 3 examined the predictive relationship between each Inner EAR scale version and PTA/WRS results. The physiologic expectation was that scores from the Inner EAR scale would be related to audiometry, since hearing-specific health arises from otologic function and scores on other scales have been associated with audiometric results.[20,21] Inner EAR scores were expected to positively correlate to WRS, as higher scores reflect better outcomes on both scales. Inner EAR scores were expected to be negatively related with PTA, as higher PTA indicates worse hearing, so we reversed PTA results in sign to ease interpretation. Standard scoring procedures were followed to recode individual responses on a scale of 0 to 100, and mean group scores were assessed.

Estimated IRT scale scores (measures of patient hearing ability) were used in the correlational analyses. Scale scores for 3-, 8-, and 11-item scales were correlated with PTA and WRS results, yielding a total of 12 correlational analyses (3 Inner EAR scale scores by 4 audiometric measures). Because univariate descriptive statistics and plots revealed considerable skewness in variable distributions, logarithmic transformations were performed to allow assessment of the strength of relationships, which is a standard approach in linear regression. Analyses were performed with Stata 14.0 (College Station, Texas).

Results

Data were collected from 353 discrete clinic visits in which patients completed the Inner EAR scale electronically after arrival. These visits represented consecutive completed responses among all new and established patients who completed the Inner EAR survey. All newly presenting patients underwent audiometry. Among the 353 patients who responded to the Inner EAR scale, 329 completed all questions, and most of the remaining patients excluded only the first item.

Analysis 1: Dimensionality of the Inner EAR Scale

Average scores across the 11-item scale demonstrated high internal consistency with a Cronbach's alpha of 0.93 (n = 353). This value supports very high reliability, similar to values obtained from large-scale educational test scores for admissions and accountability.[22,23] **Table 1** presents item means and interitem correlations, which were moderate to high (range, 0.60-0.84; median, 0.74). Overall, these results imply that the scale is internally consistent (ie, the items have a cohesive focus on hearing ability) such that scores would be unlikely to differ if similar items were used.

Factor Analysis and Principal Components Analysis. A 1-factor model was fit to the data to determine whether items

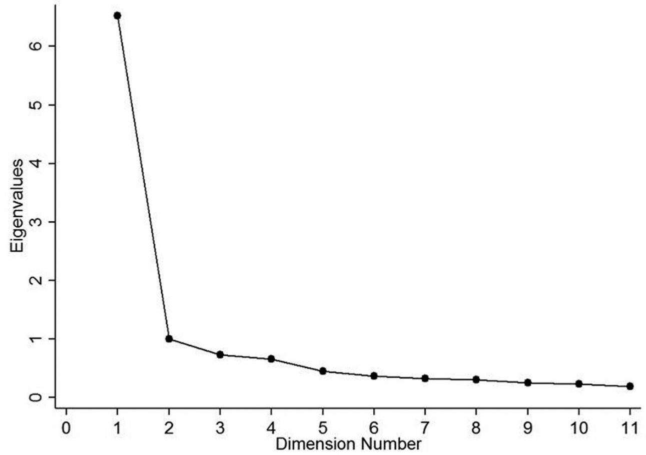

Figure 1. Scree plot of eigenvalues showing the variation accounted for by each dimension (out of 11). Estimated from a principal components analysis of standardized variables (n = 353).

measured a single underlying latent dimension. Standardized factor loadings ranged from 0.62 (ear 1, ear5b) to 0.86 (ear 3a; **Table 1**). Model fit indices were within acceptable ranges[24] (χ^2 = 335.69, *df* = 44, root mean square error of approximation = 0.14, comparative fit index = 0.88, standardized root mean square residual = 0.06), indicating that a single common factor (ie, hearing-related ability) can account for the relationships among item responses. A principal components analysis shows that an ideally weighted composite of item scores accounts for 59.4% of total variation. A scree plot from this analysis (**Figure 1**) indicates that this first component accounts for substantially more variation than subsequent composites, also suggesting that the Inner EAR scale is measuring a unidimensional ability (ie, it describes hearing-related ability substantially more so than other abilities).

Analysis 2: IRT to Estimate Item Location and Information Parameters

IRT Analysis. An IRT graded response model was fit to the full sample (n = 353) to estimate location and information parameters (Appendix 1, in the online version of the article). The information parameter estimates (a_i) ranged from 1.63 (ear11: global item) to 4.52 (ear3a: different listening situations). These results indicate that the Inner EAR scale items effectively distinguish among patients with similar but nonequal latent perceived hearing abilities. Higher values indicate greater information: the ear3a item regarding different listening situations measures perceived hearing ability better than the ear11 global item.

While information parameters describe how much information items hold, location parameters describe where on the scale this information is located. Each item has as many location parameters as it has differences among response points, so an item with 4 answer options will have 3 location parameters, while an item that allows patients to rate their symptoms on an 11-point scale will have 10 parameters. Higher location values may be preferable when distinguishing among patients with good hearing ability; lower location values may be preferable when distinguishing among patients with poor hearing ability. **Table 2** reports the estimated item location and information parameters.

These IRT parameters can be used to derive 2 useful graphs for evaluating items: category characteristic curves and item information functions. Category characteristic curves (Appendix 2, in the online version of the article)

Table 2. Item Information and Location Parameter Estimates Based on a Graded Response Model (n = 353).[a]

		Location Parameter Estimates				
Item Code	Information Parameter Estimates: a	b1	b2	b3	b4	b5
ear11 (b1-b5)	1.63	–4.24	–2.39	–1.92	–1.30	–0.70
ear11 (b6-b10)		–0.28	0.45	0.79	1.55	2.57
ear2a	3.79	–1.37	–0.10	0.94	1.88	
ear2b	2.72	–1.78	–0.65	0.62	1.62	
ear2c	2.70	–0.65	0.62	1.72	2.90	
ear2d	3.05	–0.84	0.37	1.42	2.32	
ear2e	2.33	–1.55	–0.42	0.79	1.72	
ear3a	4.52	–1.17	0.20	1.16	2.09	
ear3b	2.63	–0.97	–0.09	0.98	1.96	
ear4	2.45	–1.43	–0.42	0.79	1.41	
ear5a	1.69	–1.23	–0.05	1.59		
ear5b	1.86	–1.60	–0.73	0.10		

[a] The ear11 item has 11 score points and thus 10 location parameters that delineate adjacent score points. The ear2, ear3, and ear4 items have 5 score points and thus 4 location parameters. The ear5 items have 4 score points and thus 3 location parameters. See **Table 1** for item prompts, score means, and standard deviations.

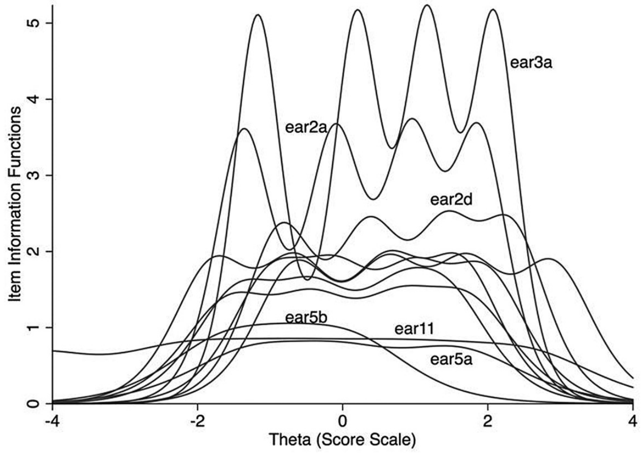

Figure 2. Item information functions estimated from a graded response model (n = 353). The 4 highest and 3 lowest item information functions are labeled. Dropping the 3 lowest leads to the 8-item scale. Keeping the 3 highest leads to the 3-item scale. The sums of these item information functions compose the test information functions shown in **Figure 3**.

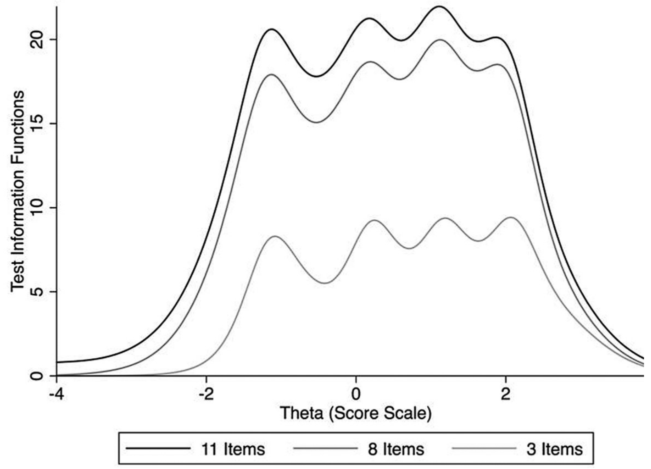

Figure 3. Test information functions estimated from a graded response model (n = 353).

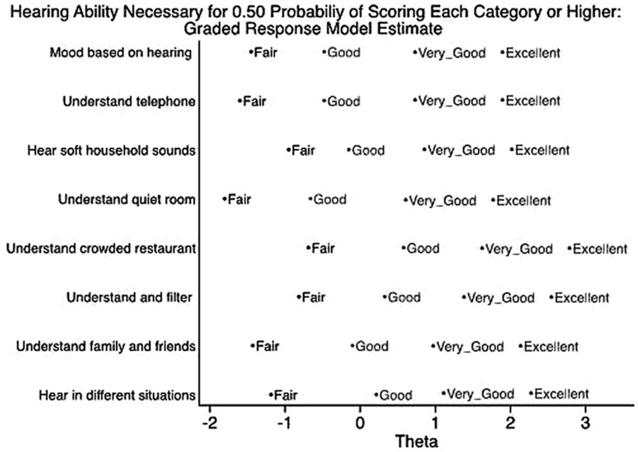

Figure 4. Item map for the 8-item scale estimated from a graded response model (n = 353).

show the probability of selecting a specific response across the range of perceived hearing ability; the height of each curve represents the probability of selecting each response. Items ear4 (mood), ear5a (need repetition), and ear5b (activity restriction) have curves with a high degree of overlap among the possible response curves. This large overlap demonstrates that these items have less distinction among responses. Specifically, there is high overlap and little distinction between ''very good'' and ''excellent'' response categories for item ear4 and between ''a little bothered'' and ''not bothered'' for items ear5a and ear5b.

The information in the presented data can be even more directly illustrated via item information functions (**Figure 2**). Information can be interpreted as precision or a reduction in random score variability. In these curves, items with greater information have higher peaks (ear3a, ear2a, and ear2d perform best). In contrast, items with flatter curves provide comparatively little information (ear5a, ear5b, and ear11 perform worst). These results suggest that the scale might function similarly with these latter 3 items removed.

A key benefit of IRT is ''local independence''—that is, that estimates of item function will not depend on features of particular patients or items. This means that we can model the information provided by shorter scales by simply subtracting the appropriate item information functions. Then, if the information provided by fewer items is similar to the information obtained by more items, it suggests that the scale may be shortened while still providing reliable and precise score estimates. **Figure 2** presents the individual item information functions, the sums of which compose the ''test information functions'' in **Figure 3** for shortened 8- and 3-item scales. **Figure 2** shows that there is balanced information across a substantial range of the scale (±2 SD). Eliminating the 3 lowest-information items reduces information but not substantially. To evaluate whether the assumption of local independence held, we refit separate IRT graded response models for the 8- and 3-item scales, and the test information results were similar.

In addition to examining test information across all values of perceived hearing ability (scale values, denoted θ [or theta]), IRT enables ''item mapping'' to provide more substantive interpretations of estimated scale scores. The core principle is that knowing a patient's θ value allows for prediction of his or her responses to items that were measured on the same scale, even if the patient does not actually respond to every item. **Figure 4** plots the items by response category locations on the θ scale (perceived hearing ability). A patient with a θ of 0 (average hearing ability in this population) is likely to answer ''good'' for most but not all items. Items focused on the ability to understand conversations in a crowded restaurant and the ability to understand and filter out unwanted noise appear to require a higher θ for patients to be likely to respond ''good'' or above.

When the 3 items with the lowest estimated information parameters were excluded to yield the 8-item scale, the Cronbach's alpha value was 0.92, negligibly less than the 0.93 reliability of the 11-item scale. The 3-item scale used ear3a (different listening situations), ear2a (understanding family and close friends), and ear2d (filter out unwanted noises), which had the highest information parameter estimates (4.52, 3.79, and 3.05, respectively). An assessment of these 3 items alone also had high internal consistency (Cronbach's alpha, 0.89). However, reducing the scale to 3 items arguably results in loss of substantive clinical value. Providers would no longer have direct information on the patient's mood related to hearing loss or perception of soft household sounds or telephone conversations, although they could predict this from item maps. Such a scale could be used when efficiency is paramount and costs of misclassification are not severe.

Analysis 3: Correlation between Inner EAR Scales and Audiometry

Correlational analysis was performed for the 55 patients who also underwent audiometric testing (**Table 3**). Scale scores from the 11-item original instrument and the proposed 8- and 3-item versions of the Inner EAR scale had statistically significant correlations with PTA and WRS results, and correlation coefficient values were moderate, ranging from 0.3 to 0.5.

These data suggest that the 8-item scale could be used in place of the 11-item scale with negligible loss of diagnostic information. As noted, reducing the scale to 3 items may arguably result in the loss of important substantive information; however, the results support using the 3-item scale in cases where efficiency is paramount.

Table 3. Correlation Coefficients for Transformed Variables in Analysis 1.[a]

	WRS′		PTA′′	
	Left (n = 54)	Right (n = 52)	Left (n = 53)	Right (n = 55)
3-item scale	0.29	0.45	0.31	0.35
95% CI	0.03-0.52	0.11-0.58	0.15-0.60	0.13-0.59
6-item scale	0.37	0.50	0.40	0.42
95% CI	0.20-0.64	0.26-0.68	0.24-0.66	0.25-0.67
8-item scale	0.40	0.48	0.42	0.43
95% CI	0.05-0.54	0.15-0.61	0.16-0.62	0.14-0.60
11-item scale	0.39	0.49	0.39	0.43
95% CI	0.10-0.57	0.18-0.62	0.19-0.63	0.18-0.62

[a]All correlations were significant at $P < .05$. WRS′: transformed word recognition scores, $x' = -\log(102 - x)$. Higher scores indicate higher levels of auditory effectiveness. PTA′′: pure tone audiometry, $x = \log(x)$. Higher scores indicate lower levels of auditory effectiveness.

Discussion

The Inner EAR scale functions well as an 11-item scale for perceived hearing ability. IRT determination of the location and information of each item provides insight into the independent information contributed by each item and supports the use of a shorter 8-item scale in routine clinical situations where increased efficiency is needed. In situations where high patient volume is anticipated so that time savings are paramount, survey burden can be decreased without losing substantial information about perceived hearing ability. If decreasing the survey by 3 questions lowers the time for completion by 1 to 2 minutes per patient, then over the course of 240 patients in clinic, 4 to 8 hours in administrative burden are saved.

IRT also provides an item map based on location and information parameters, which clinicians can use to understand how patients with a particular final score are likely to respond across individual items. For example, patients with a θ of 0 (final score, approximately 40) are most likely to have indicated "good" ability to understand in a quiet room or on the telephone but "fair" ability in a crowded restaurant or when trying to filter out background noise. Patients with a theta of –1 (final score, approximately 20) most likely indicated that their ability to hear across all items is "fair," which could prompt discussion about amplification.[25,26]

While audiometry is the gold standard, it requires access to equipment and skilled audiologists. It also accrues cost. Additional methods to screen hearing-related health thus have appeal, particularly in resource-limited environments. Recent studies assessed whether self-reported hearing loss is predictive of audiometry,[27-36] including the Blue Mountains Hearing Study, which used a single question ("Do you feel you have a hearing loss?") for evaluation, coupled with the Shortened Hearing Handicap Inventory for the Elderly.[27] These data demonstrated a negative predictive value of 96% to 98% for moderate hearing loss but a positive predictive value of just 25% to 33%. A similar result was obtained in a rural population: a sensitivity of 96% but a specificity of just 26%.[32] A systematic review of 10 longitudinal studies assessed whether 1 global question could predict audiometry, and results were mixed: sensitivity ranged from 14% to 100%, while specificity was 50% to 93%.[28] For mild but confirmed hearing loss, a given screening question may be only moderately useful,[27,37-39] and the accuracy of the self-reported screening can decrease, depending on the threshold definition for hearing impairment.[39] Here, Inner EAR scales correlate with audiometry, and further research regarding the predictive value and diagnostic properties of the shortened scales relative to the original version would be ideally be forthcoming.

Conclusions

The 11-item Inner EAR instrument demonstrates favorable characteristics of a unidimensional scale. IRT demonstrates that a shortened 8-item scale also reflects the underlying θ (perceived hearing ability), suggesting viability for use in high-volume clinical situations. In cases where further efficiency is paramount, a 3-item scale has utility for prediction of other unmeasured item responses, as well as audiometric results. As otolaryngology becomes increasingly invested in patient-reported outcomes, novel applications of advanced methods such as IRT may support scale validation and optimization of measurement efficiency.

Acknowledgments

We thank Margaret Toro and Deborah Waldman for their audiologic expertise and Andrea Fetrow for assistance with formatting, as well as Thomas Lin for additional support during the preparation of this manuscript.

Author Contributions

Annika Jessen, draft writing, substantial contributions to conception and design, data analysis and formatting, interpretation of data, revisions for intellectual content, final approval; **Andrew D. Ho,** draft writing, substantial contributions to conception and design, data analysis and formatting, interpretation of data, revisions for intellectual content, final approval; **C. Eduardo Corrales,** draft writing, substantial contributions to the interpretation of the data, revisions for intellectual content, final approval; **Bevan Yueh,** concept development and origin, draft and analysis contributions, revisions for intellectual content, final approval; **Jennifer J. Shin**, draft writing, substantial contributions to conception and design, acquisition of data, and interpretation of data, revisions for intellectual content, corresponding author, final approval.

Disclosures

Competing interests: J.J.S. receives textbook royalties from *Evidence-Based Otolaryngology* (Springer) and *Otolaryngology Prep and Practice* (Plural Publishing) and is a recipient of a Harvard Medical School Shore Foundation / Center for Faculty Development Grant and a Brigham Care Redesign Program Award.

Sponsorships: None.

Funding source: None.

Supplemental Material

Additional supporting information is available in the online version of the article.

References

1. Centers for Medicare and Medicaid Services. Medicare program. *Fed Regist*. 2016;81:79562-79892.
2. Rosenkrantz AB, Nicola GN, Allen B Jr, Hughes DR, Hirsch JA. MACRA, MIPS, and the new Medicare Quality Payment Program: an update. *J Am Coll Radiol*. 2017;14:316-323.
3. Wagle N. Implementing patient-reported outcomes. *NEJM Catalyst*. http://catalyst.nejm.org/implementing-proms-patient-reported-outcome-measures/. Published 2016.
4. Carroll TL, Lee S, Lindsay R, Locandro D, Randolph GW, Shin JJ. Evidence-based medicine in otolaryngology, part 6: patient-reported outcomes in clinical practice. *Otolaryngol Head Neck Surg*. 2018;158:8-15.
5. Shin JJ, Hartnick CJ, Randolph GW, eds. *Evidence-Based Otolaryngology*. New York, NY: Springer; 2008.
6. Edelen M, Reeve O. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res*. 2007;16(suppl 1):5-18.
7. Hays RD, Brown J, Brown LU, Spritzer KL, Crall JJ. Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of their children's dental care. *Med Care*. 2006;44(11)(suppl 3):S60-S68.
8. Lord FM. *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ: Lawrence Erlbaum Associates; 1980.
9. Lord FM. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educ Psychol Meas*. 1967;2:989-1020.
10. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing; 1968.
11. Yen WM, Fitzpatrick AR.Item response theory. In: Brennan R, ed. *Educational Measurement*. 4th ed. Westport, CT: American Council on Education/Praeger Publishers; 2006:111-118.
12. Shargorodsky J, Curhan SG, Curhan GC, et al. Change in prevalence of hearing loss in US adolescents. *JAMA*. 2010;304:772-778.
13. Walling AD, Dickson GM. Hearing loss in older adults. *Am Fam Physician*. 2012;85:1150-1156.
14. Office of Disease Prevention and Health Promotion. Healthy People 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/hearing-and-other-sensory-or-communication-disorders. Published 2010. Accessed March 14, 2017.
15. Stachler RJ, Chandrasekhar SS, Archer SM, et al. Clinical practice guideline: sudden hearing loss. *Otolaryngol Head Neck Surg*. 2012;146(3):S1-S35.
16. Alford RL, Arnos KS, Fox M, et al. American College of Medical Genetics and Genomics guideline for the clinical evaluation and etiologic diagnosis of hearing loss. *Genet Med*. 2014;16:347-355.
17. Yueh B, McDowell JA, Collins M, et al. Development and validation of the Effectiveness of [corrected] Auditory Rehabilitation scale. *Arch Otolaryngol Head Neck Surg*. 2005;131:851-856.
18. Shin JJ. An electronic interface to routinize outcomes assessment and streamline clinic workflow. *Laryngoscope*. 2017; 127:1058-1060.
19. Coster W, Ludlow L, Mancini M. Using IRT variable maps to enrich understanding of rehabilitation data. *J Outcome Meas*. 1999;3:123-133.
20. Manrique-Huarte R, Calavia D, Huarte Irujo A, et al. Treatment for hearing loss among the elderly: auditory outcomes and impact on quality of life. *Audiol Neurootol*. 2016; 21(suppl 1):29-35.
21. Maeda Y, Sugaya A, Nagayasu R, Nakagawa A, Nishizaki K. Subjective hearing-related quality-of-life is a major factor in the decision to continue using hearing aids among older persons. *Acta Otolaryngol*. 2016;136:919-922.
22. Reardon SF, Ho AD. Practical issues in estimating achievement gaps from coarsened data. *J Educ Behav Stat*. 2015;40:258-189.
23. The College Board. *Test Characteristics of the SAT: Reliability, Difficulty Levels, Completion Rates*. New York City, NY: The College Board; 2015.
24. Kline R. *Principles and Practice of Structural Equation Modeling*. 4th ed. New York, NY: The Guilford Press; 2015.
25. Collins MP, Liu CF, Taylor L, et al. Hearing aid effectiveness after aural rehabilitation: individual versus group trial results. *J Rehabil Res Dev*. 2013;50:585-598.

26. Collins MP, Souza PE, Liu CF, et al. Hearing aid effectiveness after aural rehabilitation—individual versus group (hearing) trial: RCT design and baseline characteristics. *BMC Health Serv Res*. 2009;9:233.
27. Sindhusake D, Mitchell P, Smith W, et al. Validation of self-reported hearing loss: the Blue Mountains Hearing Study. *Int J Epidemiol*. 2001;30:1371-1378.
28. Valete-Rosalino CM, Rozenfeld S. Auditory screening in the elderly: comparison between self-report and audiometry. *Braz J Otorhinolaryngol*. 2005;71:193-200.
29. Ferrite S, Santana VS, Marshall SW. Validity of self-reported hearing loss in adults: performance of three single questions. *Rev Saude Publica*. 2011;45:824-830.
30. Ramkissoon I, Cole M. Self-reported hearing difficulty versus audiometric screening in younger and older smokers and non-smokers. *J Clin Med Res*. 2011;3:183-190.
31. Ranganathan B, Counter P, Johnson I. Validation of self-reported hearing loss using television volume. *J Laryngol Otol*. 2011;125:18-21.
32. Deepthi R, Kasthuri A. Validation of the use of self-reported hearing loss and the Hearing Handicap Inventory for elderly among rural Indian elderly population. *Arch Gerontol Geriatr*. 2012;55:762-767.
33. Diao M, Sun J, Jiang T, et al. Comparison between self-reported hearing and measured hearing thresholds of the elderly in China. *Ear Hear*. 2014;35:e228-e232.
34. Kamil RJ, Genther DJ, Lin FR. Factors associated with the accuracy of subjective assessments of hearing impairment. *Ear Hear*. 2015;36:164-167.
35. Brennan-Jones CG, Taljaard DS, Brennan-Jones SE, et al. Self-reported hearing loss and manual audiometry: a rural versus urban comparison. *Aust J Rural Health*. 2016;24:130-135.
36. Manchaiah V. Role of self-reported hearing disability and measured hearing sensitivity in understanding participation restrictions and health-related quality of life: a study with hundred and three older adults with hearing loss. *Clin Otolaryngol*. 2017;42:924-926.
37. Clark K, Sowers M, Wallace RB, et al. The accuracy of self-reported hearing loss in women aged 60-85 years. *Am J Epidemiol*. 1991;134:704-708.
38. Nondahl DM, Cruickshanks KJ, Wiley TL, et al. Accuracy of self-reported hearing loss. *Audiology*. 1998;37:295-301.
39. Bagai A, Thavendiranathan P, Detsky AS. Does this patient have hearing impairment? *JAMA*. 2006;295:416-428.