Article:
Hornsby, B.W.Y., Camarata, S., Cho, S.-J. et al. (3 more authors) (2021) Development and Validation of the Vanderbilt Fatigue Scale for Adults (VFS-A). Psychological Assessment, 33 (8). pp. 777-788. ISSN: 1040-3590

https://doi.org/10.1037/pas0001021

Published in final edited form as:

Psychol Assess. 2021 August ; 33(8): 777–788. doi:10.1037/pas0001021.

Development and Validation of the Vanderbilt Fatigue Scale for Adults (VFS-A)

Benjamin W.Y. Hornsby[1], **Stephen Camarata**[1], **Sun-Joo Cho**[2], **Hilary Davis**[1], **Ronan McGarrigle**[3], **Fred H. Bess**[1]

[1]Vanderbilt University School of Medicine, Department of Hearing & Speech Sciences, Nashville, TN, United States

[2]Vanderbilt University, Peabody College, Department of Psychology & Human Development, Nashville, TN, United States

[3]Department of Psychology, University of York, UK

Abstract

Listening-related fatigue can be a significant burden for adults with hearing loss (AHL), and those with other health or language-related issues (e.g., multiple sclerosis, traumatic brain injury, second language learners) who must allocate substantial cognitive resources to the process of listening. The 40-item Vanderbilt Fatigue Scale for Adults (VFS-A-40) was designed to measure listening-related fatigue in AHL and other populations. This paper describes the development, and psychometric properties, of the VFS-A-40. Initial qualitative analyses in AHL suggested listening-related fatigue was multidimensional, with physical, mental, emotional, and social domains. However, exploratory factor analyses revealed a unidimensional structure. Item and test characteristics were evaluated using Item Response Theory (IRT). Results confirmed that all test items were of high quality. IRT analyses revealed high marginal reliability and an analysis of test-retest scores revealed adequate reliability. In addition, an analysis of differential item functioning provided evidence of good construct validity across age, gender, and hearing loss groups. In sum, the VFS-A-40 is a reliable and valid tool for quantifying listening-related fatigue in adults. We believe the VFS-A-40 will be useful for identifying those most at risk for severe listening-related fatigue and for assessing interventions to reduce its negative effects.

Keywords

Vanderbilt Fatigue Scale; listening-related fatigue; fatigue; adults with hearing loss; outcome measures

Address correspondence to Benjamin W.Y. Hornsby, Department of Hearing & Speech Sciences, Vanderbilt University School of Medicine, Vanderbilt Bill Wilkerson Center, 1215 21st Avenue South, MCE South Tower, Room 8310 Nashville, Tennessee 37232-8718, United States. ben.hornsby@vumc.org.

The authors have no conflicts of interest to disclose.

Introduction

Fatigue is commonly defined as a subjective experience, or mood, that is associated with feelings of weariness, tiredness, a lack of vigor or energy, or decreased motivation to continue a task (Hornsby et al., 2016). Feelings of fatigue are a common consequence of sustained and demanding physical or mental effort and something everyone experiences at some point in their lives. In a healthy population these feelings tend to be mild, transient in nature, and resolve with a short rest or break from the demanding activity. This type of fatigue is normal and healthy as it encourages us to evaluate the costs and benefits of the current fatiguing activity (Hockey, 2013). However, for some people their feelings of fatigue are more severe and occur more frequently, often developing in response to normal daily activities such as routine house cleaning, self-care, or even simply listening in a noisy setting. For these individuals', fatigue can be debilitating and have a significant negative effect on quality of life (Bess & Hornsby, 2014; Davis et al., 2020; Evans & Wickstrom, 1999; Robinson-Smith et al., 2000).

Many subjective measures have been developed to quantify the frequency and severity of fatigue in adults. Some instruments quantify fatigue as part of a global assessment of general health or mood. The Profile of Mood States (McNair et al., 1971) is an example of one such measure. The POMS assesses fatigue, vigor, and several other mood states, including depression, tension, anger, and confusion. Other tools focus more specifically on the assessment of fatigue as a unidimensional or multidimensional construct. Some measures, like the POMS fatigue subscale, are generic and are appropriate for measuring fatigue associated with a wide range of chronic health conditions. Other scales are designed to assess fatigue associated with a specific disease state (e.g., cancer) or population (e.g., athletes). For example, the Revised Piper Fatigue Scale (Piper et al., 1998) was designed to assess cancer-related fatigue. Both generic and disease-specific measures are useful; however, well-designed, disease-specific, instruments are generally more sensitive than generic measures, at least for their targeted population (Patrick & Deyo, 1989).

In this paper, our focus is on *listening-related fatigue*, a type of fatigue that can result from the sustained application of mental effort while listening (Davis et al., 2020; Pichora-Fuller et al., 2016) and on the development of a reliable and valid scale to measure this construct. Listening-related fatigue may be especially problematic for people with hearing loss (see Hornsby et al., 2016, for review); the third most common chronic health condition in the United States (Masterson, 2016). For example, recent data from focus groups and interviews suggest that some people with hearing loss experience high levels of effort and stress while trying to listen and communicate, thus increasing their risk for listening-related fatigue (Davis et al., 2020; Holman et al., 2019). In addition, researchers using generic fatigue scales, or scales developed for other populations, have reported increased fatigue in adults and children with hearing loss (Alhanbali et al., 2017; Hornsby et al., 2016; Hornsby, 2013; Hornsby et al., 2017; Hornsby & Kipp, 2016). However, these findings are not universal.

For example, Hornsby and Kipp (2016) found no difference in *mean* POMS fatigue scores between a group of adults seeking help for hearing difficulties and age-matched normative data. There were, however, significant differences in the prevalence of *severe*

fatigue between groups (defined as POMS fatigue scores that were >1.5 standard deviations above normative ratings). A similar finding, also using the POMS, was observed by Dwyer et al. (2019). They compared mean fatigue ratings between a group of college-age adults with moderate-to-profound hearing loss and age-matched peers without hearing loss, and found no differences between groups. However, in the same study, Dwyer et al. (2019) asked participants about their *listening-related fatigue* using study-specific questions (e.g., "Difficulty listening causes me to become physically or emotionally tired"). Results using these listening-specific items showed adults with hearing loss (AHL) experienced significantly more listening-related fatigue than their age-matched cohort (effect size r ranged from .69 to .86 across items). These findings suggest that generic fatigue measures may not be sensitive to the fatigue commonly experienced by adults and children with hearing loss. In addition, current clinical assessment procedures do not include direct measures of effort and fatigue at all and standard assessments (e.g., pure tone threshold or speech testing) are not sensitive to these constructs (Hornsby & Kipp, 2016; Alhanbali et al., 2017). A reliable and valid measure of listening-related fatigue would help to provide clinicians with a more holistic and detailed picture of an individual's specific communication challenges. The lack of such a measure motivated the current study.

Importantly, the negative effects of listening-related fatigue are not limited to those with hearing loss. Listening is a complex sensory and cognitive skill that, when disrupted, can lead to significant fatigue in other populations. For example, fatigue resulting from sustained cognitive effort on auditory (or visual) tasks is a common complaint in persons with multiple sclerosis (Bryant et al., 2004; Paul et al., 1998) and traumatic brain injury (TBI; Johansson et al., 2009). Likewise, even in healthy populations, those who must exert additional effort when processing speech, like non-native speakers and second-language learners, may also suffer from increased listening-related fatigue (Borghini & Hazan, 2018; Kellerman, 1992). Research in these related areas, coupled with the ubiquitous role that listening plays in our society (e.g., in academic, vocational, and social arenas), highlights the potentially far-reaching impact of listening-related fatigue and the need for a reliable measurement tool.

The purpose of the current study was to develop the Vanderbilt Fatigue Scale for Adults (VFS-A). Item Response Theory (IRT) was used to validate the scale. IRT is a measurement model that relates an individual's response to test items to an underlying latent trait (e.g., Embretson & Reise, 2000). In this case, the underlying latent trait, or theta (θ), is listening-related fatigue. IRT has been shown to be effective for developing and optimizing the sensitivity of such scales. The VFS-A was designed to assess listening-related fatigue in AHL but may also be appropriate for adults with other auditory or communication difficulties (e.g., learning disabilities, auditory processing disorders, second language learners).

Methods

Scale development and validation of the VFS-A was divided into four phases (See Figure 1). Briefly, in Phase 1, we conducted a literature review and held focus groups with AHL to operationalize the latent construct of listening-related fatigue. In Phase 2, we used Phase 1 results to guide development of a large pool of potential test items which were reviewed for

clarity, relevance, and redundancy. A subset of high-quality items was selected for further analyses. In Phase 3.1, we collected data from adults with and without hearing loss (*N*=580 respondents) using this reduced item pool. These data were used to evaluate item quality and select smaller pool of high-quality items for review by an expert panel (Phase 3.2). Using an iterative process (described below) we identified 40, high-quality, items for a 40-item version of the VFS-A (the VFS-A-40). In Phase 4, we conducted analyses to describe the psychometric properties of the VFS-A-40. We also assessed test-retest reliability of the VFS-A-40 in a group of 86 adult cochlear implant users. A detailed description of each phase is provided below. All study procedures were reviewed and approved by the Vanderbilt Institutional Review Board (IRB#140946).

Participants and Data Collection

Across all phases, study participants were recruited locally from Vanderbilt Bill Wilkerson Center (VBWC) Audiology clinics and from the surrounding community. In addition, we used a variety of online postings to recruit broadly. Phase 1 focus group participants included 43 adults (32 female), between 20 and 77 years of age (mean/median 53.5/54 years; standard deviation/semi-interquartile range 16.2/12 years) with mild to profound hearing loss. Focus group comments were used to inform scale content and aid in development of scale items (see Davis et al., 2020, for details). In Phase 2, we conducted cognitive interviews with a small group of AHL (*N*=8; 7 female) to evaluate test items for comprehension, clarity, and content relevance. Interviewees were adults between 23 and 75 years of age (mean/median 53.9/55 years; standard deviation/semi-interquartile range 14.9/5.4) with moderate-to-severe bilateral hearing loss.

Our Phase 3.1 sample included adults (*N* = 580) aged 18–88 years (mean/median = 50.2/52 years; standard deviation/semi-interquartile range = 16/13 years). Most respondents reported their ethnicity as non-Hispanic (89%; 8.4% did not report), their race as white (90.5%; 4.7% did not report) and were female (73.4%; 6.4% did not report). Hearing status was determined in response to the question: "Do you have a hearing loss?" Approximately 75% (*n* = 433) reported having a hearing loss. Of those with self-reported hearing loss, approximately 74% (*n* = 322) reported using a hearing device (e.g., hearing aid or cochlear implant). In addition, those with a self-reported hearing loss rated their degree of hearing impairment using, a 5-point scale, as none (*n* = 6), mild/slight (*n* = 71), moderate (*n* = 159), severe (*n* = 102), or profound (*n* = 95). In Phase 3.2 our expert panel (*N*=11) consisted of adults with (*n*=4) and without hearing loss (*n*=7) who were (a) research scientists with expertise in various aspects of hearing loss, (b) active, or retired, clinical audiologists or (c) both. All panel members were familiar with the psychosocial consequences of hearing loss either through their work, their own personal experiences having a hearing loss, or both.

Finally, in Phase 4, we assessed the test-retest reliability of a preliminary version of the VFS-A-40[1] in a separate sample of eighty-six (41 female/1 did not report) adult cochlear

[1]These test-retest data were collected with an early version of the VFS-A-40 which included an item that was only appropriate for someone using a hearing device (e.g., a hearing aid, cochlear implant). This item ("I need to remove or turn off my hearing device to take a break from listening."), was replaced in a later version. The replacement item ("I get headaches after taking part in group conversations.") had similar psychometric properties but was not specific to device usage.

implant users. Participants were aged 24–88 years old (mean/median=63.5/66.5 years; standard deviation/semi-interquartile range = 14.7/9.4 years) with at least 12 months of experience with their current cochlear implant(s). In addition, all participants reported no changes to their implants in the 3 months prior to completing the initial (baseline) survey and no changes to their implants in the time between completing their initial and follow up surveys. Fifty-one of the participants had a single implant while 35 participants wore bilateral implants.

Item and Test Development Methods

Phase 1 Methods: Latent Construct Operationalization—In Phase 1, we conducted a review of the broader fatigue literature and an in-depth review of literature focusing on the relationship between fatigue and hearing loss. Review results suggested the construct of fatigue was complex and its relation to hearing loss was understudied (see Hornsby et al., 2016 for review). We then recruited adults (N=43) with a wide range of hearing losses to participate in focus groups to discuss their understanding and experiences with listening-related fatigue. Focus group discussions were recorded, transcribed, coded, and analyzed to identify the most common themes related to listening-related fatigue from the perspective of AHL (for details of this process see Davis et al., 2020). We used information from these focus groups to develop a construct map (Wilson, 2005) that operationalized listening-related fatigue in terms of hypothesized domains and severity levels. We used this construct map to guide the development of potential items for the VFS-A (see Results section for details).

Phase 2 Methods: Item Creation and Initial Assessment

Phase 2.1 Methods: Initial Item Development and Item Reduction Process.: In this phase we used information from focus groups to inform development of specific items for the proposed VFS-A. Specifically, at least two research team members were assigned to review each participant comment from each focus group transcript. If a comment focused on behaviors, feelings, and/or situations relevant to listening-related fatigue, team members created a potential test item based on the relevant comment. During this initial phase, multiple versions of a scale item could be created from a single participant comment.

This large initial pool of items was then reviewed by the research team to identify a subset of clear and relevant items and remove less clear, redundant items. The process was iterative in nature with separate team members identifying items they felt best assessed a given aspect of listening-related fatigue (e.g., severe, emotional). All potential items were coded based on their hypothesized fatigue domain and severity level, as described in our construct map. For example, the item “I spend a lot of energy just trying to listen.” was hypothesized to assess severe listening-related fatigue in the cognitive domain. This item was one of multiple items targeting severe cognitive fatigue and was thus coded as “cognitive_3_2” (where Domain = Cognitive; Severity level = 3-Severe; and Item number = 2; each domain/severity level had multiple items).

Phase 2.2 Methods: Initial Item Assessment via Cognitive Interviews.: Next, we used cognitive interviews to identify interpretation or comprehension problems with our

preliminary test items. In this process eight participants were asked to read an item and to "think out loud" as they made an answer. The "think out loud" responses and rationales provided insight into how the respondents processed each question (Collins, 2003). Follow-up questions were used to ensure that the respondent's interpretation of a test item matched our original intent. Interviews were recorded and transcribed for later evaluation.

Phase 3 Methods: Item Assessment, Reduction, and Development of the VFS-A-40

Phase 3.1 Methods: EFA and IRT analyses.: In Phase 3, we collected responses using our pool of potential test items from a large sample of adults with and without hearing loss. We then used exploratory factor analysis (EFA) to examine the latent structure of our item pool and IRT to examine the characteristics of the test items in the pool. The goal of these analyses was to identify high-quality items for our scale. An overview of the analysis methods is provided below.

EFA Methods.: In Phase 3 we first investigated the latent structure (i.e., the number of dimensions and factor loading patterns) of the VFS-A 106 item pool by conducting a series of EFA using Mplus Version 8.3 (Muthen & Muthen, 1998–2017). Using polychoric correlations (specifically, weighted least square with adjusted means and variance [WLSMV] with Oblimin rotation and Oblique type) a series of EFAs were conducted, extracting 1–4 factors. Fit indices were compared across 1–4 factor models. We used the following empirically supported guidelines to assess the goodness of model fit: a root-mean-square error of approximation index (RMSEA; Steiger & Lind, 1980) of < .06, a root-mean-square residual (RMSR) of < .08, a comparative fit index (CFI; Bentler, 1990) and Tucker-Lewis index (TTL; Tucker & Lewis, 1973) > .95 (Hu & Bentler, 1999; Yu, 2002).

IRT Analyses Methods.: We evaluated individual item quality, and the quality of sets of items, using IRT analyses. Based on EFA results, a unidimensional graded response model (GRM; Samejima, 1969) or a multidimensional GRM (De Ayala, 1994) was used to investigate the item characteristics. The GRM is an item response model for ordered polytomous responses and has two kinds of item parameters— an item discrimination parameter and item threshold parameters[2]. Item parameter estimates, individual item information, and test information were obtained using a (marginal) maximum likelihood estimation (MLE) method in Mplus. For IRT scoring, expected a posteriori (EAP) for a specific response pattern was used. Missingness in item responses was considered missing at random and treated as missing under the MLE.

[2]The item discrimination parameter is a measure of the item's ability to discriminate between various levels of the construct (e.g., differentiate between individuals with varying fatigue severity). A larger value indicates the item is sensitive to variations in the latent construct. The second kind of parameter, the item threshold parameters for each item are used to model response scores. For our data set having 5 response scores, 4-item threshold parameters are estimated: Threshold 1 reflects the transition point from a score of 0 (e.g., Never/Almost Never) to scores 1–4 (Rarely, Sometimes, Often, or Always/Almost Always); Threshold 2 reflects the transition point from scores of 0 or 1 to scores 2– 4; Threshold 3 reflects the transition point from scores of 0–2 to scores of 3–4; and Threshold 4 reflects the transition point from scores of 0–3 to a score of 4.

Item quality was evaluated based on three criteria: (a) Are item threshold estimates of GRM in order and well-separated? Ordered and well-separated item thresholds indicate that the item response anchors behaved as designed (e.g., as the latent construct increased, respondents would select higher response options). (b) Do average item locations (i.e., averaged across item threshold estimates) match their hypothesized severity level (mild, moderate, or severe) based on the construct map, and (c) Is an item discrimination estimate positive and high in magnitude? High discrimination estimates suggest an item will be able to effectively differentiate between people with varying degrees of listening-related fatigue. These item characteristics are important as they will impact the sensitivity of the final scale.

Finally, we looked at *item* and *test* information for various sets of high-quality items to help select items for the final VFS-A. *Item information* describes the amount of information *an item* provides across different levels of the latent construct (θ). *Test information* is the sum of item information across a set of items, also as a function of θ. High test information implies good measurement fidelity. We evaluated item and test information for several sets of items. Our goal was to generate a set of items that provide a test information level of at least 11.11 (test information = 1/[standard error of an IRT scorê2]) over a wide range of severities of listening-related fatigue (i.e., a range of θ's). A standard error of 0.3 was used as an empirical cut-off to calculate the target test information value of 11.11. This error value corresponds to a reliability coefficient of 0.95, which has been deemed acceptable in the development of other clinical scales (e.g., Cole et al., 2012; Hospers et al., 2016).

Phase 3.2 Methods: Expert Panel Review.: We used the IRT analyses described above to identify a subset of high-quality items for further review by a panel of experts (N=11). The panel reviewed items for relevance, clarity and completeness. Reviewers were first provided with a copy of the construct map to help them understand the construct of listening-related fatigue. They were then asked to review items for relevance and clarity and to provide comments regarding the overall comprehensiveness of the item pool. An item was defined as relevant if it reflected, sampled, and measured the construct of listening-related fatigue as described in the construct map. Relevance was rated on a four-point scale (0, 1, 2, or 3) with response options including not relevant, somewhat relevant, quite relevant, and highly relevant. An item was defined as having good clarity if it was perceived as well-written, distinct, and at an appropriate reading level for AHL. Clarity was assessed by responding yes or no to the question "Is this question well-written and easy to understand?" When considering comprehensiveness, we asked panel members to use the construct map as a guide and comment via free response on whether the item pool provided a comprehensive overview of issues important to listening-related fatigue.

Phase 4 Methods: Final Analyses of the VFS-A-40—Using feedback from the expert panel and results from IRT analyses, a final pool of 40 items was identified. Once identified, we repeated the EFA and IRT analyses described in Phase 3.1 on this final item pool. In addition to item parameter estimates and item and test information, item fit was examined to judge how well the GRM described each test item. Item fit was assessed with the generalized χ^2 test (e.g., Kang & Chen, 2008).

We also conducted differential item functioning (DIF) analyses to determine whether the final pool of test items measured listening-related fatigue equivalently across distinct populations (i.e., age, gender, and hearing loss groups). DIF analyses were implemented using lordif package (Choi et al., 2011) in R version 3.2.4 (R Core Team, 2016). An ordinal logistic regression model, in conjunction with IRT scale scores as a matching criterion was chosen to detect DIF items. For each item, DIF was evaluated assuming a uniform effect (the effect is constant across trait levels) and non-uniform effect (the effect varies across trait levels).[3] In addition to the likelihood ratio test at alpha=0.01, McFadden's pseudo R^2 measure, which is a proportional reduction in the −2 log–likelihood statistic, was chosen as a DIF effect size measure. Zumbo (1999) suggests guidelines for classifying DIF based on the pseudo R^2 statistic as negligible (< 0.13), moderate (between 0.13 and 0.26), and large (> 0.26).

As evidence of reliability for the final 40-item pool, marginal IRT reliability (Green et al., 1984) was assessed. In addition, test-retest reliability was assessed in a separate sample of experienced adult cochlear implant users (N=86). These participants were recruited as a control group for an ongoing separate study examining how listening-related fatigue changes over time following receipt of a cochlear implant. Participants completed a preliminary version of the VFS-A-40[1] twice— once to provide a baseline score and a second time approximately 3 months later (mean/median = 3.2/3.1 months; standard deviation/semi-interquartile range=0.4/0.11 months).

Results

Item and Test Development Results

Phase 1 Results: Latent Construct Operationalization—An analysis of focus group transcripts revealed that the construct of listening-related fatigue was multidimensional and centered around four primary domains: physical, cognitive, emotional, and social. Likewise, the experience of listening-related fatigue within these domains varied from mild to severe in nature. An individual's listening-related fatigue was impacted by multiple factors including the acoustic environment of the listener, their motivation to listen, and any coping strategies used to minimize fatigue-related negative effects (see Davis et al., 2020, for details). We used these focus group data to create a construct map to operationalize listening-related fatigue (see Supplementary File 1).

This construct map defined the various experiences of listening-related fatigue across domains and levels of severity. Experiences in each domain were described in terms of the feelings and behaviors that were commonly reported by focus group participants. Regarding severity of listening-related fatigue, our analysis of focus group transcripts revealed large

[3]The DIF detection was made by comparing three nested ordinal logistic regression models: (a) **Model 1**: the cumulative probability that the actual item response falls in category k or higher = intercept + slope1 * latent variable, (b) **Model 2**: the cumulative probability that the actual item response falls in category k or higher = intercept + slope1 * latent variable + slope2 * group, and (c) **Model 3**: the cumulative probability that the actual item response falls in category k or higher = intercept + slope1 * latent variable + slope2 * group + slope3 * latent variable * group. *Uniform DIF* was tested by comparing the log likelihood values for Models 1 and 2 (one degree of freedom, or df = 1) and *non-uniform DIF* by comparing Models 2 and 3 (df = 1). A *total DIF* effect was evaluated by comparing Models 1 and 3 (df = 2). For these three comparisons, twice the difference in log likelihoods was compared to a χ^2 distribution with a specified df. Type I error rate, alpha=0.01, was chosen.

individual variations. For example, an individual suffering from severe cognitive fatigue might report an inability or unwillingness to remain attentive and focused when listening—even for relatively short periods of time in a good (e.g., low noise) listening situation. A common coping strategy for this individual may be to disengage or give up trying to understand in that listening condition or avoid going into it at all. In contrast, another individual may report only minimal problems with attention or focus while listening, unless the acoustics are very challenging and/or the listening demands are extended over a long period of time. This individual's listening-related fatigue (in the cognitive domain) would be considered mild in severity. Similar feelings, behaviors, and situations were identified for the other domains. In the next Phase, we developed potential test items that tapped into the feelings and behaviors described in the construct map.

Phase 2 Results: Item Creation and Initial Assessment

Phase 2.1 Results: Initial Item Development and Item Reduction Process.: The number of test items required for a sensitive and reliable test depends, in part, on the information provided by the specific test items. However, prior research has shown that with high-quality items acceptable test precision for a given domain can be obtained with at least 10 items (Sinharay, 2010). Thus, our goal was to initially develop enough items to identify a minimum of 10 high-quality items per domain (i.e., 40 items total). However, to achieve this goal a larger item pool would be required.

The research team used transcripts from Phase 1 focus groups to guide the development of over 2,000 potential test items. A subset of the study authors (N=3) met to review these items, exclude/revise poorly worded or redundant items, and identify a smaller pool of high-quality test items for additional review. This process resulted in a subset of 302 items which were subjected to further review by the full research team. Specifically, each team member evaluated all 302 items and, individually, selected approximately 100 items that they viewed as highest quality. Team members were instructed to select items (a) to ensure coverage of all domain/severity levels and (b) based on the item's readability, clarity, and uniqueness. The research team then met as a group to review the individual selections and reach consensus on a reduced pool of items. This iterative review resulted in a reduced item pool containing 110 high-quality items that were then subjected to further evaluation via cognitive interviews.

Phase 2.2 Results: Initial Item Assessment via Cognitive Interviews.: Cognitive interviews revealed that most items were clearly understood. Interviewee responses to items were consistent with the item's underlying intent. However, based on interview feedback, several items were discarded (n = 4) or modified (n = 18) to improve clarity, resulting in a final pool of 106 potential test items. The domain and severity level of the modified items matched those of the original items.

For most items, response options were made using either (a) a 5-point Likert frequency anchor where response options included Never/Almost Never, Rarely, Sometimes, Often, and Always/Almost Always or (b) a 5-point Likert agreement anchor, where response options included Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly

Agree. One item focused on "fatigability", asking "In a difficult listening situation, I become tired after listening for ____". Response options for this item were on a 5-point scale including: less than 30 minutes, 30–60 minutes, 2–3 hours, more than 3 hours and NA- I don't become tired from difficult listening. This reduced pool of 106 items had an average Flesch-Kincaid reading grade level of 6.7 (range 0.5–11.2). These items were used to collect data for further evaluation of item quality in Phase 3 (see below).

Phase 3 Results: Item Assessment, Reduction and Development of the VFS-A-40

Phase 3.1 Results: Initial EFA and Item Assessment using a GRM.: In Phase 3 we collected data, using this pool of 106 items, from adults with and without hearing loss (N = 580; see "Participants and Data Collection" section above). We conducted preliminary analyses of their responses using EFA and a GRM. The purpose of these initial analyses was to: (a) investigate the underlying latent structure of listening-related fatigue as measured using our test items, and (b) evaluate the quality of individual items to assist in selecting the final items for the VFS-A.

Despite focus group results suggesting that listening-related fatigue was multidimensional, results from the initial EFA revealed almost all items loaded highly onto a single factor—consistent with a unidimensional model (see Table 1). Specifically, an EFA using all 106 items showed that a one-factor (unidimensional) model provided a good fit to the data according to model-data fit indices: RMSEA (0.057), RMSR (0.056), CFI (0.958), and TLI (0.958). See Supplementary file 2 for additional analyses confirming a one-factor solution. There were two items (one from the physical and social domains [P2_7 and S2_8]; both targeting a moderate severity) which did not load on this same factor. Based on this finding these two items were excluded from further analyses.

Based on the EFA results, a unidimensional GRM was fit to the remaining 104-item data set. We first conducted IRT analyses to examine the item parameter estimates of the GRM. The item parameter estimates are on the logit scale. Item discrimination estimates ranged from 1.45 to 5.8 (mean = 3.17), which are considered high magnitude item discrimination values. For almost all items, thresholds were in order and well separated. However, two items ("Listening in background noise can bother me" and "I prefer to listen in small (versus large) groups of people") showed poor threshold separation at low levels of listening-related fatigue. For these items, threshold scores were similar for ratings of 0–1 and thus excluded from consideration for use in the final scale. GRM results also revealed that several test items were more sensitive to milder (n=18) or more severe (n=6) listening-related fatigue than expected. For example, the item "Struggling to listen and understand makes me tired." was originally hypothesized to target listening-related fatigue of moderate severity. However, IRT analyses revealed the item was more sensitive to mild levels of listening-related fatigue. Discovering that an item targeted more, or less, severe fatigue than expected was not cause for exclusion. Rather, the item was recoded to accurately reflect the target severity. This new coding was used to ensure the items used in the final scale adequately covered all domains and severity levels of listening-related fatigue as defined by the construct map.

Our next step was to reduce our item pool given our goal of approximately 10 items/domain in our final scale. IRT analyses were used to identify 61 high-quality (high item information and good threshold coverage) items from the 104-item pool for further analyses. In addition to item information, items were selected to ensure adequate coverage of all domains and severity levels described in the construct map. Fifteen items were selected to target each domain (Physical[4], Social, Emotional and Cognitive) and one item assessed "fatigability" (i.e., the time it takes to become fatigued due to difficult listening). Within each domain there were more items targeting moderate (5 items) and severe (8 items) listening-related fatigue and fewer items targeting mild fatigue (2 items). We chose this approach based on the assumption that the negative effects of *mild* fatigue would be minimal. In contrast, the psychosocial and functional consequences of *moderate-to-severe* listening-related fatigue are more likely to warrant intervention. A scale sensitive to variations in moderate-to-severe listening-related fatigue would also be required to detect the benefits of any such intervention. Based on the iterative evaluation process described above, the 61 items chosen for further evaluation were viewed as clearly written and assessed relatively unique aspects of listening-related fatigue (i.e., redundant items were excluded). These items were then subjected to review by an expert panel as described below.

Phase 3.2 Results: Expert Panel Review.: Expert panel members rated the relevance and clarity of the 61 items and the comprehensiveness of the item pool in relation to the construct map. All items were rated as "quite relevant" or "highly relevant" by at least 50% of panel members. Using a 0 (Not relevant) to 3 (Highly relevant) scale, the mean relevance rating for all items was 2.3 (i.e., midway between quite relevant and highly relevant). The median rating was 3.0 suggesting most reviewers felt all items were highly relevant. In terms of clarity, all items were rated as well-written and easy to understand by at least 50% of panel members.

Panel member ratings and comments, in conjunction with results of IRT analyses, were used to select a final pool of 40 items for the VFS-A-40. To ensure the final scale adequately assessed all relevant areas of listening-related fatigue, the final item pool contained 10 items in each domain. This number of items was required to meet our empirical criterion of test information of $\geq$1.11 (standard error of 0.3), over a wide range of fatigue severities, in each domain. To ensure sensitivity to moderate-to-severe listening-related fatigue in each domain, items were chosen to target moderate (3 items) and severe (6 items) listening-related fatigue more frequently than mild fatigue (1 item). Specifically, individual item information curves were evaluated to identify items that provided high information across a wide range of fatigue severities. The test information (sum of item information over items) resulting from various sets of items were compared to identify an optimal set of items for each domain.

Thirty-seven of the final 40 items were rated as quite or highly relevant by 70–100% of panel members. The mean/median relevance rating for all 40 items was 2.7/3, respectively, again suggesting that most reviewers felt the items were highly relevant. In terms of clarity,

[4]Note after some initial analyses we replaced one item within the physical domain which asked about hearing device usage with an item that did not require respondents to use a hearing device to answer. The replacement item targeted the same domain and severity level and had similar psychometric properties as the original item.

all selected items were rated as well-written and easy to understand by at least 50% of panel members and most items (30 of 40) were viewed as well-written by 90–100% of panel members. An additional five items were described as well written and easy to understand by 80–89% of respondents. Of the remaining five items only two were rated as well written and easy to understand by fewer than 60–70% of panel members. Based on this review process, five (5) of the 40 selected items were modified slightly to improve clarity. Modifications involved adding or removing one or two words in a sentence. For example, the item "Struggling to listen and understand makes me tired." was revised to read "Struggling to listen and understand makes me *feel* tired." The single word added is shown in italics. In the next section we confirm the structure and quality of this final version of the VFS-A-40.

Phase 4 Results: Final Analyses of the VFS-A-40

Phase 4.1 Results: EFA and IRT Analyses.: The final version of the VFS-A-40 provides a total score based on all 40 items. In addition, 10-item subscale scores are also provided for each targeted domain of listening-related fatigue (physical, cognitive, social, and emotional). The mean/median Flesch-Kincaid grade reading level for the VFS-A-40 is 6.9/6.7 (standard deviation= 1.9; range = 2.8–10.3). Using data from Phase 3 respondents, we repeated our EFA and IRT analyses using only responses from this subset of 40 items.

EFA Results.: EFA model fit indices on this reduced data set continued to suggest that a one-factor model provided the most parsimonious, statistically compelling fit (See Table 2 and Supplementary file 2).

IRT Analyses Results.: Based on these EFA results, a unidimensional GRM was fit to the 40-item set data. Results revealed that item discrimination estimates remained high, with values falling between 2.110 and 5.883 (Mean=3.836). Thresholds were in order and well separated for all items. Figure 2 shows examples of category characteristic curves for four selected items- one from each domain of the VFS-A-40. Item parameter discrimination and threshold estimates of the GRM for items used to calculate a total score and subscale scores, are provided as supplementary files (See Supplementary files 3 and 4, respectively). All items fit well to the data based on the generalized χ^2 test.

Figure 3 shows test information curves (TICs) for each domain of the VFS-A-40 and for the whole test. TIC's show test information as a function of θ (i.e., the level of the underlying construct). A high value of test information and a broad TIC implies good measurement fidelity across a wide range of listening-related fatigue severities. These TICs show that the VFS-A-40 has good fidelity (test information $\geq$11.11) for the people with θ's ranging between approximately −1.5 to +1.5. When estimating listening-related fatigue using the total score, the range is even wider. The TIC based on all 40 items shows good fidelity (test information $\geq$11.11) for θ values within the range of approximately −2 to +2.2.

DIF Analyses for Construct Validity.: Following selection of items for the VFS-A-40, DIF analyses were conducted to determine whether the VFS-A-40 measured listening-related fatigue equivalently in distinct populations. We assessed DIF of the test items across age groups in two ways. First, we divided our sample into two groups based on the median

age (52 years) of our sample (i.e., group 1: age < 52; group 2: age ≥52). Second, we used a cut-point of 65 years to broadly assess DIF among working age and retired adults. In addition, we examined DIF across gender groups (self-reported as male or female) and based on self-reported hearing loss (those who answered yes or no to the question "Do you have a hearing loss?").

A detailed listing of all DIF detection results are provided in Supplementary file 5. Briefly, items were flagged as a DIF item when any of the likelihood ratio χ^2statistics were significant. There were 3, 7, and 14 DIF items based on the three likelihood ratio χ^2 statistics for age, gender, and hearing loss groups, respectively. However, the DIF effect sizes (Mc-Fadden's pseudo R^2 measure) for these items were negligible (all effect sizes were ≤0.023). Results analyzing age effects using a cut point of 65 years of age replicated those using the median age (results not shown). In this analysis there were four DIF items (c3_5, p1_1, p3_4, and s3_1). However, their DIF effect sizes were again negligible (< 0.012). Based on these results, we conclude VFS-A-40 scores can be interpreted in the same way between younger and older adults, males and females, and those with and without self-reported hearing loss. Taken together, the results of these analyses suggest the VFS-A-40 has acceptable evidence of construct validity.

Test Score Reliability.: To assess test score reliability, we evaluated marginal IRT reliability and test-retest reliability. Marginal IRT reliability (ranging from 0 to 1) of the final scale version was evaluated using data from Phase 3 participants (N=580). The marginal IRT reliability of the final version of the VFS-A-40 was high at .981. Marginal IRT reliability for cognitive (.948), emotional (.949), physical (.942), and social (.941) domains was also high, providing evidence of good test score reliability. Test-retest reliability was assessed using the VFS-A-40[1] in a group of experienced adult cochlear implant users (N=86) who were participating in a separate, ongoing, study. A series of Wilcoxon signed ranks tests were used to examine differences in mean total, and subscale, scores obtained at baseline (T1) and approximately 3-months later (T2). Results revealed no significant difference between T1 and T2 total scores or any T1 and T2 subscale score. Mean T1-T2 differences in summed scores were all < 1 point. In addition, we assessed temporal stability by examining correlations, using Pearson correlation coefficients (i.e., test-retest reliability coefficients) and ICC's, between individual T1 and T2 total and subscale scores. All correlation coefficients were positive and statistically significant, ranging from .60 to .69 (see Table 3). Relative to other generic and disease-specific fatigue measures and existing measures of other transient states (see our summary and discussion section), we believe these results suggest the VFS-A-40 has adequate test-retest stability.

SUMMARY AND DISCUSSION

In this paper we describe our process for developing the VFS-A-40, a scale designed to assess listening-related fatigue in adults. The final version of the scale is provided in the supplementary materials (See Supplementary file 6). We first used data obtained in focus groups from AHL to operationalize the construct of listening-related fatigue. Results from these groups suggested that listening-related fatigue was a complex construct with physical, cognitive, emotional and social manifestations. For some people, the functional

and psychosocial consequences of listening-related fatigue can be significant. For example, someone with severe listening-related fatigue may need additional sleep or rest following a challenging listening situation or be unable to maintain focus and attention while listening in that setting. In addition, individuals may become extremely sad or upset due to their listening difficulties and/or isolate themselves from social settings that involve sustained and/or challenging listening. These findings highlight the potential negative effects that listening-related fatigue can have on the quality of life of those most affected.

We used comments from focus group participants to develop a pool of potential test items and then evaluated the psychometric properties of those items using responses from a large (N=580) and diverse sample of adults with and without hearing loss. The pool contained items reflecting all domains and severity levels of listening-related fatigue as identified by focus group participants. Despite this targeted approach, an EFA revealed that almost all items loaded heavily onto a single, primary, factor. This result suggests that the diverse clinical expressions of listening-related fatigue are related, and reflective of a single underlying construct; a finding that is consistent with research in the broader fatigue literature. For example, Michielsen et al. (2004) argued that the multidimensional nature of some existing generic and disease-specific fatigue scales may be overestimated due to the statistical approach used to examine dimensionality (i.e., EFA with eigenvalues exceeding unity as a criterion for identifying unique factors). These authors reexamined the dimensionality of several existing, multidimensional, scales using an EFA and Mokken Scale analysis. In contrast to prior work, their results suggested all scales were unidimensional in nature. In addition, an EFA on responses from all the measures combined showed that all items loaded heavily onto a single factor. They argue this finding offers strong support for the view of fatigue as a unidimensional construct. Lai et al. (2006) reported a similar finding when exploring the dimensionality of cancer-related fatigue.

Given our findings from adults with and without hearing loss, we believe the total score from the VFS-A-40 provides the most precise, robust, and psychometrically sound measure of an individual's listening-related fatigue. However, examining subscale scores may still be clinically useful despite associations between domains. For example, consider an intervention designed to improve an individual's emotional responses to listening-related fatigue. In this case we might predict the largest effects on items targeting the emotional domain. This effect could, in theory, be identified by an analysis of subscale scores. Using the total score alone, however, could potentially mask the benefits of such an intervention. Moreover, subscale scores may have clinical utility by identifying intervention priorities. For example, a patient's fatigue score may indicate moderate listening-related fatigue across domains. However, follow-up questioning reveals that their primary concern is feeling *physically* exhausted after group meetings. In this case, the clinician may find responses from the physical subscale useful for generating intervention priorities and patient counseling. In addition, while results from our sample of adults with and without hearing loss suggest listening-related fatigue is a unidimensional construct, this may not be the case in other populations. For example, a recent study by McGarrigle et al. (2020) used the preliminary version of the VFS-A-40[1] to examine effortful listening and fatigue in young and older adults with relatively good hearing. Exploratory analyses revealed significant age-related differences in listening-related fatigue— but only in the social domain. No group

differences were observed in total scores or in any other domain (McGarrigle et al., 2020). For these reasons we have maintained subscale scoring options in our final version of the VFS-A-40.

Importantly, IRT analyses confirmed the good reliability of the VFS-A-40 total score (based on all 40 items) and subscale scores (based on 10 items/domain). An analysis of test information for the total score and subscale scores showed good fidelity (test information ≥11.11; IRT marginal reliability coefficient ≥.95) over a wide range of severity levels (see Figure 3). Test information, and thus the precision of the IRT scores, was reduced for individuals experiencing very mild ($\theta < -2$) or more severe ($\theta > 2.2$) listening-related fatigue. A similar pattern was seen for subscale scores with test precision somewhat reduced for those with lower ($\theta < -1.5$) and higher levels ($\theta > 1.5$) of domain-specific, listening-related fatigue. It is also noteworthy that the vast majority of our respondents (~94%) had total scores that fell within θ's of −2 to +2.2 and only ~1% of respondents (7 of 580) had θ's >2.2, suggesting the VFS-A-40 can provide a precise estimate of listening-related fatigue for most adults (see Figure 4). A similar finding was observed for subscale scores with ~82% to 88% of respondents subscale scores falling within θ's ranging from approximately −1.7 to +1.9. Theta values where test information fell below 11.11 varied slightly across domains from a minimum of −1.7 (cognitive domain) to a maximum of 1.9 (physical domain; results not shown).

Our analyses also suggest the VFS-A-40 has adequate test-retest reliability. There are no universal "cut-off" scores for reliability as it is inherently dependent on the construct being measured (e.g., a stable trait or a variable state) and the sensitivity of the scores (Thompson, 2002). Listening-related fatigue, like the broader construct of general fatigue and other moods, is a variable state that can be impacted by many factors. Thus, we expect estimates of listening-related fatigue to vary over time, impacting measures of test-retest reliability. Test-retest variability measured over a 3-month period revealed correlation coefficients ranging from .60–.69 across the VFS-A-40 scales. These values are consistent with the reliability of other generic and disease-specific fatigue measures, and existing measures of other transient states (Krueger & Schkade, 2008; McNair & Heuchert, 2010; Donovan et al., 2015).

For example, Krueger and Schkade (2008) examined reliability of several measures of subjective well-being, an important but variable state. When assessing well-being over a 2 to 2.5-month period, reliability coefficients ranged from .50 – .82. In comparison to another fatigue measure, Donovan et al. (2015) found reliability coefficients across several studies that used the Multidimensional Fatigue Symptom Inventory-Short Form (Stein et al., 2004) ranged from .51 to .70. Likewise, test-retest reliability of the fatigue subscale of the POMS, a widely used generic fatigue measure, varied from .39 to ~.75 across studies (Gibson, 1997; McNair & Heuchert, 2010; Salinsky et al., 2001).

Despite our positive findings, additional work is needed to further evaluate the reliability and validity of VFS-A-40. For example, literacy is an important issue for scale development and ensuring usability. Our initial work using cognitive interviews and an expert panel suggests the items comprising the VFS-A-40 are clear, comprehensible, and accessible to adults

with hearing loss. Likewise, personal experience in our lab using the final version of the VFS-A-40 on ongoing projects suggests most adults can complete the scale, on their own, in 5–10 minutes. However, additional work specifically examining usability and accessibility of the scale for those with varying literacy levels is needed.

In addition, this initial work targeted adults with and without self-reported hearing loss. However, we believe listening-related fatigue may be an important problem for other populations that struggle with listening-related difficulties (e.g., second language learners, tinnitus sufferers, or those with additional learning or other auditory processing disorders, etc.). For example, recent work from our laboratory using a generic fatigue measure revealed that school-age children with hearing loss who were also poor readers reported more cognitive fatigue than children with hearing loss who were good readers (Camarata et al., 2018). Associations between listening-related fatigue and other academic or learning difficulties are unknown but warrant additional investigation.

Finally, additional studies are needed to examine relationships between the VFS-A-40 and other general, and hearing-loss specific, subjective, behavioral, and clinical outcomes (e.g., depression, social isolation, self-reported hearing difficulties, willingness to seek help for hearing difficulties, etc.). Such studies are essential for determining the functional impact of listening-related fatigue and establishing clinical criteria for intervention and for assessing intervention benefits (i.e., minimal clinically important differences).

Regarding clinical utility, we used IRT analyses to evaluate the psychometric properties of the VFS-A-40. Specifically, we used EAP to analyze item response patterns rather than summed scores, in part, because of its sensitivity to differences in the underlying latent construct among participants. However, calculating a VFS-A-40 summed score (total and subscale scores) provides a simple alternative for clinical use when IRT analysis of response patterns is not possible. Therefore, we also employed EAP on respondent summed scores (Lord & Wingersky, 1984) and have provided conversion tables to relate total and subscale summed scores to IRT scale scores (see Supplementary files 7 and 8, respectively). The EAP for a given summed score is calculated as an average IRT score over all possible response patterns. In addition, we have provided R-code which uses item discrimination and threshold estimates derived from this study to calculate IRT scores based on VFS-A-40 response data. The R code and item parameter estimates have been made freely available for download online (https://osf.io/dpy9m/).

In summary, despite its impact on diverse populations (e.g., AHL, adults with TBI and other cognitive disorders, second language learners), until now a measure of fatigue specific to listening and communication challenges did not exist. In this paper, we provide a detailed description of the development of a novel scale for measuring listening-related fatigue in adults. Standard analytic assessment criteria reveal that the VFS-A-40 is a reliable and valid measure of the underlying unidimensional construct of listening-related fatigue. The VFS-A-40 will help to identify those most affected by listening-related fatigue and will bolster the current measurement toolkit of clinicians and researchers. A more comprehensive understanding of the negative impact of listening difficulties will ultimately help to tailor intervention strategies (e.g., use of a hearing device) that seek to improve quality of life.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by a grant from Starkey Inc (Hornsby, PI), by the NIH NIDCD Grant #R21DC012865 (Hornsby, PI), by an NICHD Grant P30HD15052 to the Vanderbilt Kennedy Center for Research on Human Development, and by a Vanderbilt Institute for Clinical and Translational Research grant (UL1 TR000445 from NCATS/NIH). The opinions expressed are those of the authors and do not represent the views of the NIH or other institutions. Portions of this work were presented at the 2018 American Academy of Audiology meeting, Nashville, Tennessee. R-code for calculating VFS-A-40 IRT (Item response theory) scale scores is available at https://osf.io/dpy9m/.

REFERENCES

Alhanbali S, Dawes P, Lloyd S, & Munro KJ (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. Ear & Hearing, 38(1), e39–e48. [PubMed: 27541332]

Bentler PM (1990). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 107, 238–246. 10.1037/0033-2909.107.2.238 [PubMed: 2320703]

Bess FH, & Hornsby BW (2014). Commentary: Listening can be exhausting—Fatigue in children and adults with hearing loss. Ear & Hearing, 35(6), 592. [PubMed: 25255399]

Borghini G, & Hazan V (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. Frontiers in neuroscience, 12, 152. [PubMed: 29593489]

Bryant D, Chiaravalloti ND, & DeLuca J (2004). Objective Measurement of Cognitive Fatigue in Multiple Sclerosis. Rehabilitation psychology, 49(2), 114.

Camarata S, Werfel K, Davis T, Hornsby BW, & Bess FH (2018). Language abilities, phonological awareness, reading skills, and subjective fatigue in school-age children with mild to moderate hearing loss. Exceptional Children, 84(4), 420–436.

Choi SW, Gibbons LE, & Crane PK (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. Journal of Statistical Software, 39(8), 1.

Cole DA, Cho S-J, Martin NC, Youngstrom EA, March JS, Findling RL, Compas BE, Goodyer IM, Rohde P, Weissman M, Essex MJ, Hyde JS, Curry JF, Forehand R, Slattery MJ, Felton JW, & Maxwell MA (2012). Are increased weight and appetite useful indicators of depression in children and adolescents? Journal of Abnormal Psychology, 121(4), 838–851. 10.1037/a0028175 [PubMed: 22686866]

Collins D (2003). Pretesting survey instruments: an overview of cognitive methods. Quality of Life Research, 12(3), 229–238. [PubMed: 12769135]

Davis H, Schlundt D, Bonnet K, Camarata S, Bess FH, & Hornsby B (2020). Understanding Listening-Related Fatigue: Perspectives of Adults with Hearing Loss. International Journal of Audiology, 1–11. 10.1080/14992027.2020.1834631

De Ayala R (1994). The influence of multidimensionality on the graded response model. Applied Psychological Measurement, 18(2), 155–170.

Donovan KA, Stein KD, Lee M, Leach CR, Ilozumba O, & Jacobsen PB (2015). Systematic review of the multidimensional fatigue symptom inventory-short form. Supportive Care in Cancer, 23(1), 191–212. [PubMed: 25142703]

Dwyer RT, Gifford RH, Bess FH, Dorman M, Spahr A, & Hornsby BWY (2019). Diurnal Cortisol Levels and Subjective Ratings of Effort and Fatigue in Adult Cochlear Implant Users: A Pilot Study. American Journal of Audiology, 28(3), 686–696. 10.1044/2019_AJA-19-0009 [PubMed: 31430174]

Embretson SE, & Reise SP (2000). Item response theory for psychologists. Lawrence-Erlbahum.

Evans EJ, & Wickstrom B (1999). Subjective fatigue and self-care in individuals with chronic illness. Medsurg Nursing, 8(6), 363. [PubMed: 11000775]

Gibson SJ (1997). The measurement of mood states in older adults. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 52(4), P167–P174.

Green B, Bock RD, Humphreys LG, Linn RL, & Reckase MD (1984). Technical Guidelines for Assessing Computerized Adaptive Tests Journal of Educational Measurement, 21(4), 347–360. 10.1111/j.1745-3984.1984.tb01039.x

Hockey R (2013). The psychology of fatigue: Work, effort and control. Cambridge University Press. 10.1017/CBO9781139015394

Holman JA, Drummond A, Hughes SE, & Naylor G (2019). Hearing impairment and daily-life fatigue: a qualitative study. International Journal of Audiology, 58(7), 408–416. 10.1080/14992027.2019.1597284 [PubMed: 31032678]

Hornsby BW (2013). The Effects of Hearing Aid Use on Listening Effort and Mental Fatigue Associated With Sustained Speech Processing Demands. Ear & Hearing, 34(5), 523–534. [PubMed: 23426091]

Hornsby BW, Naylor G, & Bess FH (2016). A Taxonomy of Fatigue Concepts and Their Relation to Hearing Loss. Ear & Hearing, 37 Suppl 1, 136S–144S. 10.1097/AUD.0000000000000289 [PubMed: 27355763]

Hornsby BWY, Gustafson SJ, Lancaster H, Cho S-J, Camarata S, & Bess FH (2017). Subjective Fatigue in Children With Hearing Loss Assessed Using Self- and Parent-Proxy Report. American Journal of Audiology, 26(3S), 393–407. 10.1044/2017_AJA-17-0007 [PubMed: 29049623]

Hornsby BWY, & Kipp AM (2016). Subjective Ratings of Fatigue and Vigor in Adults With Hearing Loss Are Driven by Perceived Hearing Difficulties Not Degree of Hearing Loss. Ear and Hearing, 37(1), e1–e10. 10.1097/aud.0000000000000203 [PubMed: 26295606]

Hospers JMB, Smits N, Smits C, Stam M, Terwee CB, & Kramer SE (2016). Reevaluation of the Amsterdam Inventory for Auditory Disability and Handicap Using Item Response Theory. Journal of Speech, Language, and Hearing Research, 59(2), 373–383. 10.1044/2015_JSLHR-H-15-0156

Hu L, & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55. 10.1080/10705519909540118

Johansson B, Berglund P, & Rönnbäck L (2009). Mental fatigue and impaired information processing after mild and moderate traumatic brain injury. Brain injury, 23(13–14), 1027–1040. [PubMed: 19909051]

Kang T, & Chen TT (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. Journal of Educational Measurement, 45(4), 391–406.

Kellerman S (1992). 'I see what you mean': The role of kinesic behaviour in listening and implications for foreign and second language learning. Applied linguistics, 13(3), 239–258.

Krueger AB, & Schkade DA (2008). The reliability of subjective well-being measures. Journal of public economics, 92(8–9), 1833–1845. [PubMed: 19649136]

Lai J-S, Crane PK, & Cella D (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Quality of Life Research, 15(7), 1179. [PubMed: 17001438]

Lord FM, & Wingersky MS (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". Applied Psychological Measurement, 8(4), 453–461.

Masterson EA (2016). Hearing impairment among noise-exposed workers—United States, 2003–2012. MMWR. Morbidity and mortality weekly report, 65.

McGarrigle R, Knight S, Rakusen L, Geller J, & Mattys S (in press). Older adults show a more sustained pattern of effortful listening than young adults. Psychology & Aging.

McNair D, Lorr M, & Droppleman L (1971). Profile of mood states. Educational and Industrial Testing Service. http://www.mhs.com/product.aspx?gr=cli&id=overview&prod=poms

McNair DM, & Heuchert JWP (2010). Profile of mood states: Technical update. Multi-Health Systems Inc.

Michielsen HJ, De Vries J, Van Heck GL, Van de Vijver FJ, & Sijtsma K (2004). Examination of the dimensionality of fatigue. European Journal of Psychological Assessment, 20(1), 39–48.

Muthen L & Muthen B (1998-2017). Mplus User's Guide (8th Edition). Muthen & Muthen.

Patrick DL, & Deyo RA (1989). Generic and disease-specific measures in assessing health status and quality of life. Medical care, S217–S232. [PubMed: 2646490]

Paul RH, Beatty WW, Schneider R, Blanco CR, & Hames KA (1998). Cognitive and physical fatigue in multiple sclerosis: relations between self-report and objective performance. Applied neuropsychology, 5(3), 143–148. [PubMed: 16318452]

Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BW, Humes LE, Lemke U, Lunner T, Matthen M, & Mackersie CL (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). Ear and Hearing, 37, 5S–27S. [PubMed: 27355771]

Piper BF, Dibble SL, Dodd MJ, Weiss MC, Slaughter RE, & Paul SM (1998). The revised Piper Fatigue Scale: psychometric evaluation in women with breast cancer. Oncol Nurs Forum, 25(4), 677–684. https://www.ncbi.nlm.nih.gov/pubmed/9599351 [PubMed: 9599351]

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Robinson-Smith G, Johnston MV, & Allen J (2000). Self-care self-efficacy, quality of life, and depression after stroke. Archives of physical medicine and rehabilitation, 81(4), 460–464. [PubMed: 10768536]

Salinsky MC, Storzbach D, Dodrill CB, & Binder LM (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12–16-week period. Journal of International Neuropsychological Society, 7(5), 597–605. 10.1017/s1355617701755075

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34(1), 97. 10.1007/BF03372160

Sinharay S (2010). How often do subscores have added value? Results from operational and simulated data. Journal of Educational Measurement, 47(2), 150–174.

Steiger JH, & Lind J (1980). Statistically-based tests for the number of common factors. Paper presented at the Annual Spring Meeting of the Psychometric Society, IA.

Stein KD, Jacobsen PB, Blanchard CM, & Thors C (2004). Further validation of the multidimensional fatigue symptom inventory-short form. Journal of pain and symptom management, 27(1), 14–23. [PubMed: 14711465]

Thompson B (2002). Score reliability: Contemporary thinking on reliability issues. Sage publications.

Tucker LR, & Lewis C (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1–10. 10.1007/BF02291170

Wilson M (2005). An item response modeling approach. Lawrence Erlbaum Associates.

Yu CY (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Unpublished doctoral dissertation). University of California, Los Angeles, CA.

Zumbo BD (1999). A handbook on the theory and methods of differential item functioning (DIF). Ottawa: National Defense Headquarters, 1–57.

Public Significance Statement:

Listening-related fatigue can negatively affect people with a wide range of health conditions, including hearing loss. A reliable and valid assessment method is required to identify those with moderate-to-severe fatigue who may need interventions, and to assess the efficacy and effectiveness of any such interventions. The 40-item Vanderbilt Fatigue Scale for Adults (VFS-A-40) was developed to fill these needs.

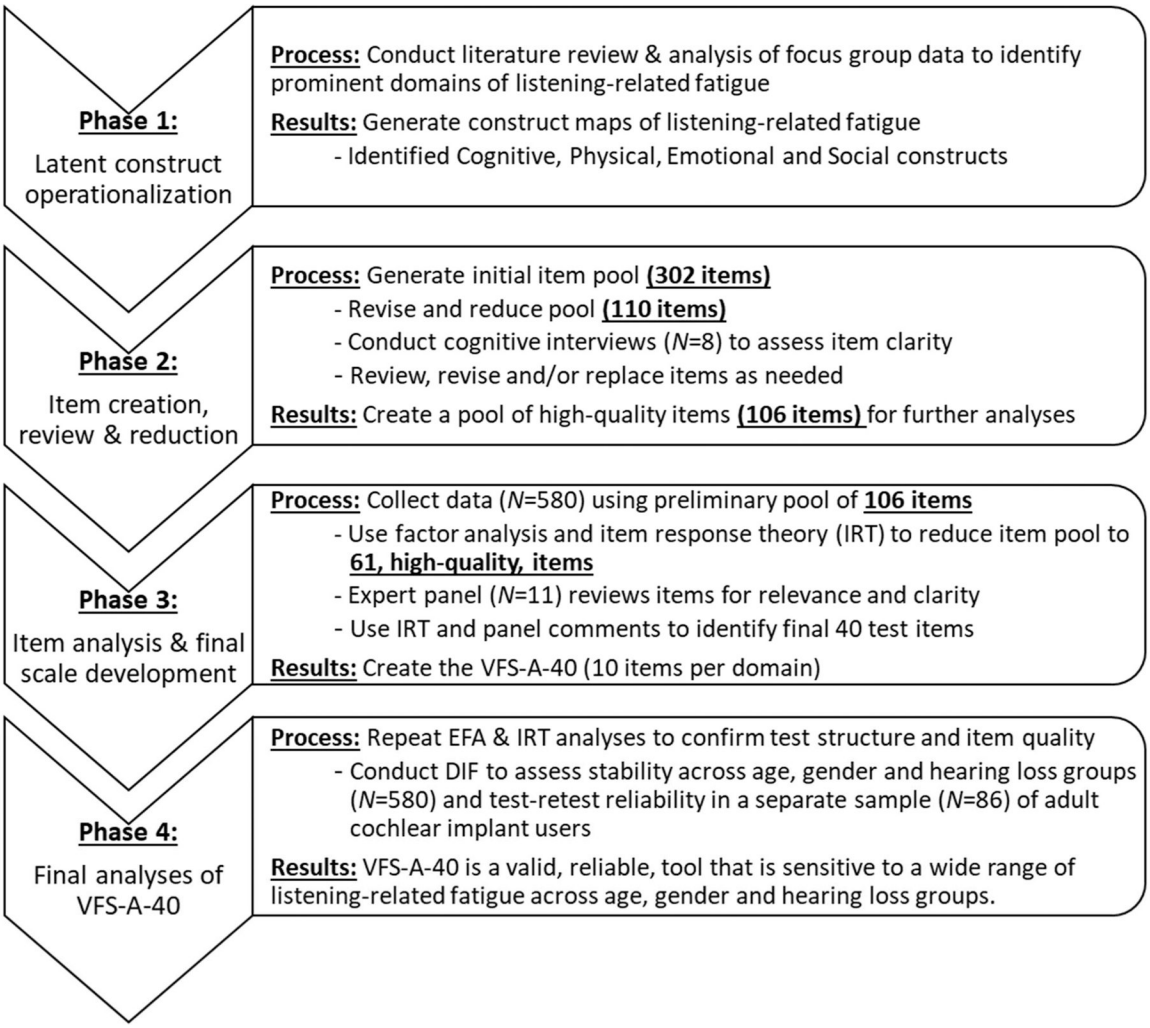

Figure 1.
Flow chart of the VFS-A scale development and validation process

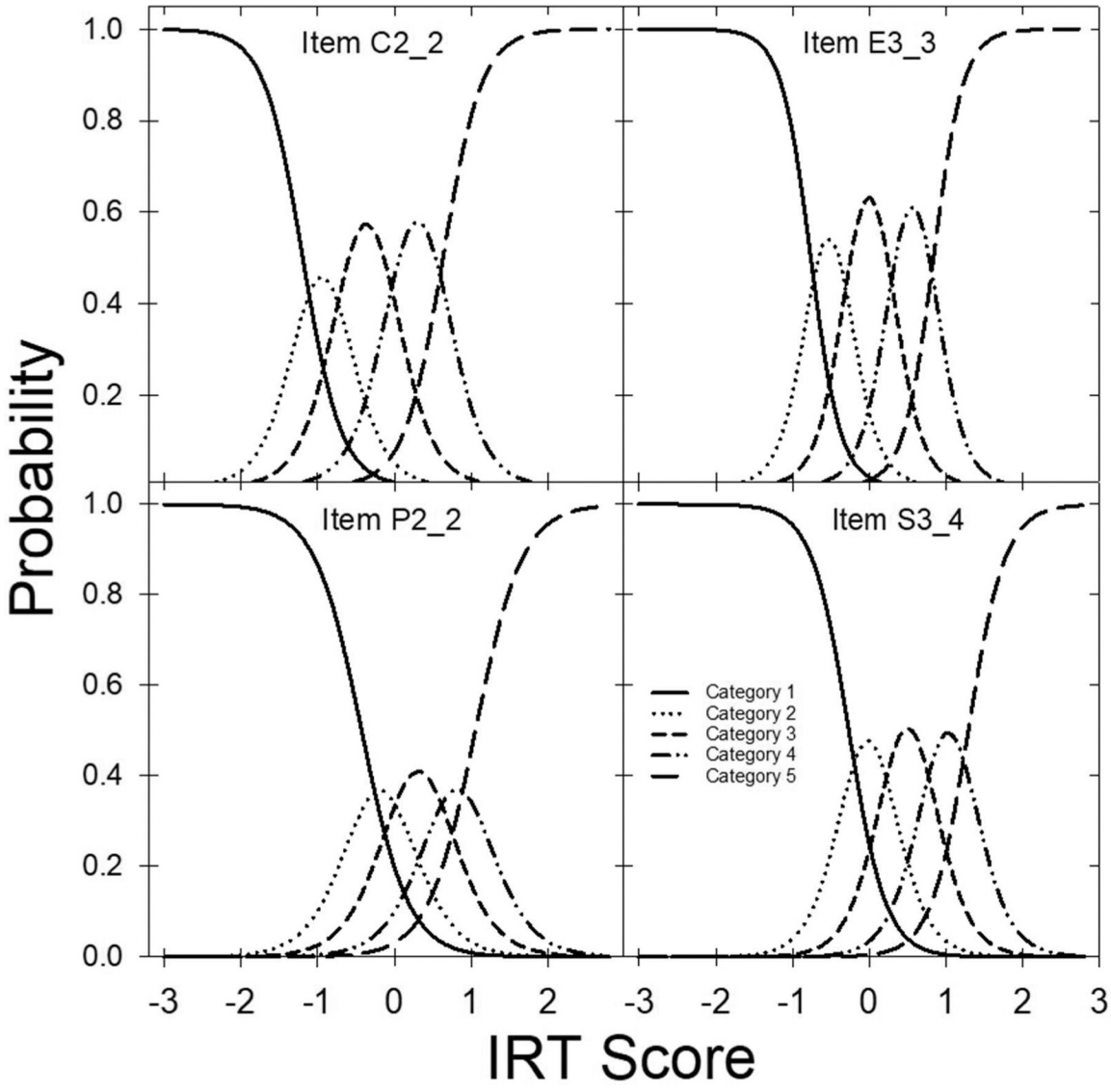

Figure 2. Exemplar category characteristic curves for four items from the VFS-A-40.

Note. IRT scores are shown on the x-axis and the probability of selecting a given response option (e.g., Category 1=Never/Almost Never) is shown on the y-axis. Panels show a category characteristic curve for a test item from the Cognitive (Item C2_2), Emotional (Item E3_3), Physical (Item P2_2) and Social (Item S3_4) domains, respectively

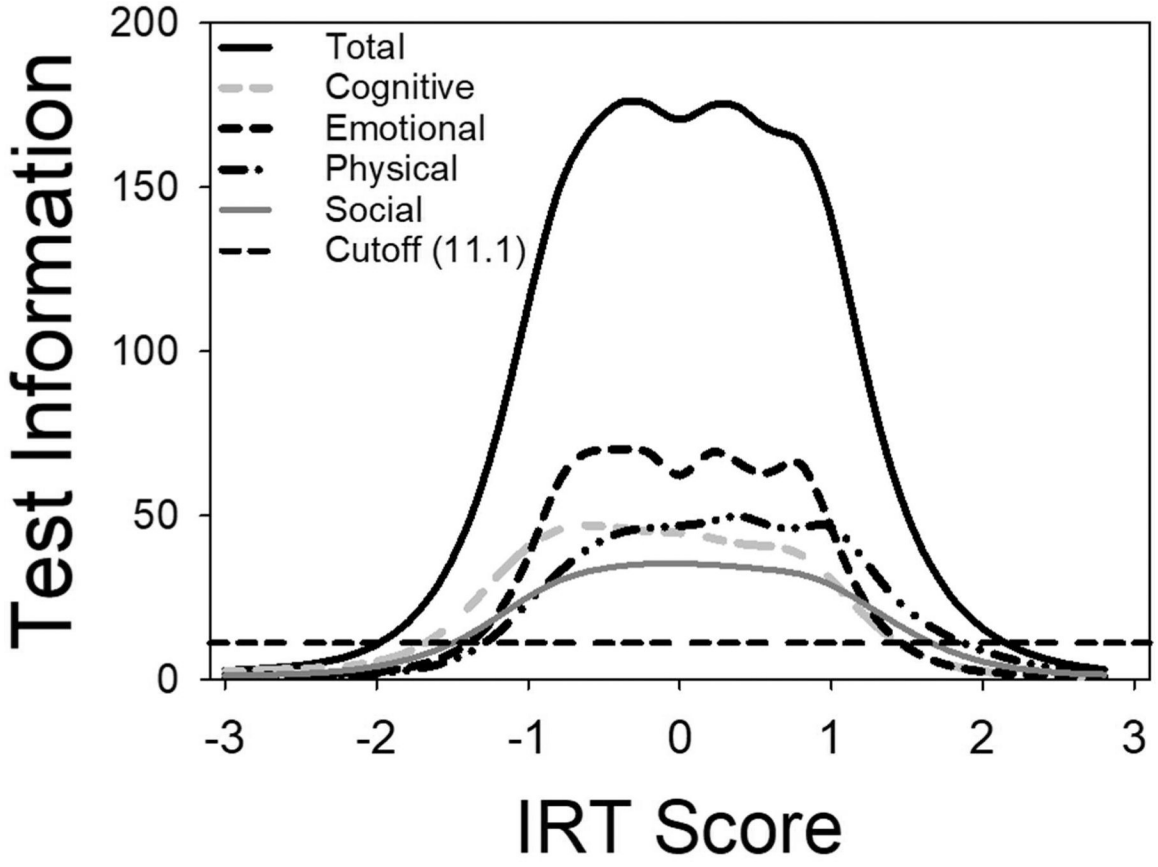

Figure 3. VFS-A-40 Test information curves (TIC)

Note. TIC's for the VFS-A-40 total test and the Cognitive, Physical, Emotional and Social subscales. The dashed line parallel to the *x*-axis represents a test information level of 11.1which corresponds to a reliability coefficient of 0.95. The intersection of the TICs and dashed line represents the range of scores over which the VFS-A-40 demonstrates of acceptable test information.

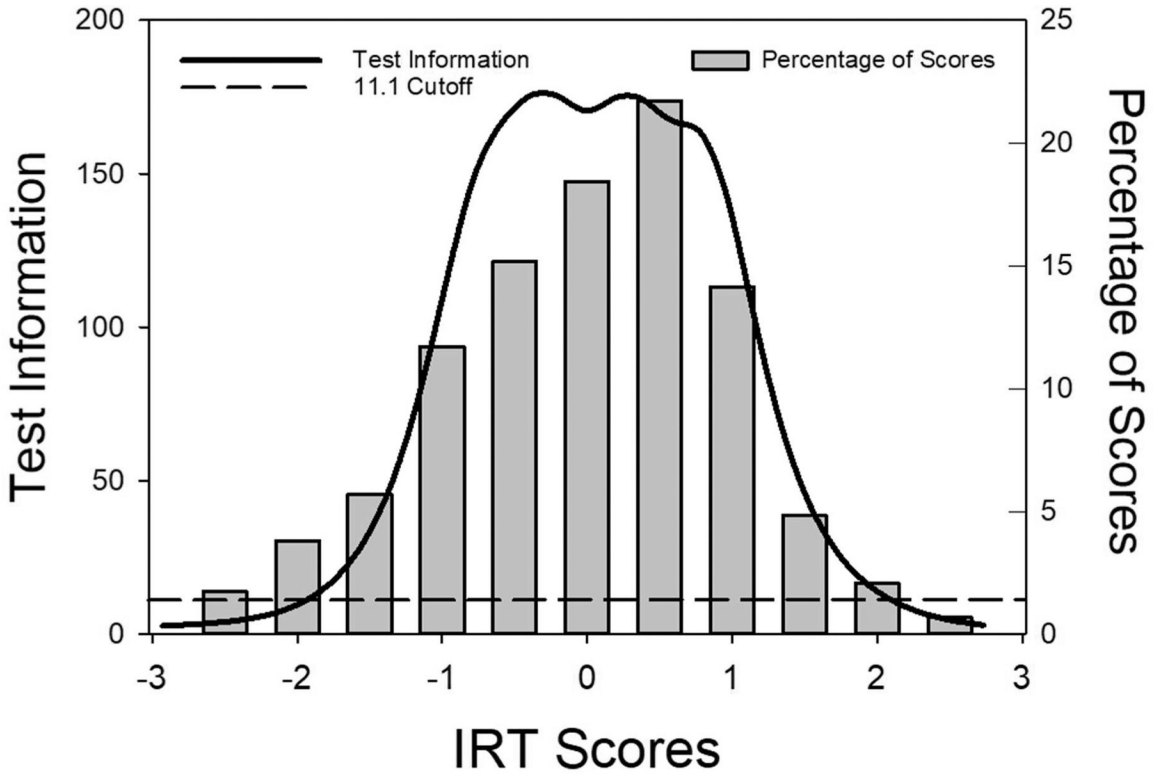

Figure 4. Distribution of IRT scores in our Sample in Relation to Test Information of the VFS-A-40

Note. Distribution of total IRT scores for 580 respondents (Grey bars; right *y*-axis represents percentage of respondent IRT scores falling within a certain range) and test information of the VFS-A-40 (solid line; left *y*-axis represents test information) as a function of IRT score (x-axis). The dashed line represents test information of 11.1 which corresponds to a reliability coefficient of .95.

Table 1.

EFA results based on 106-item pool used to develop the VFS-A-40.

	Fit Indices across 1–4 factor EFA			
Fit Indices	**1-Factor**	**2-Factor**	**3-Factor**	**4-Factor**
RMSEA	0.057[0.056,0.058]	0.054[0.053,0.055]	0.049[0.047,0.050]	0.044[0.043,0.045]
SRMR	0.056	0.040	0.035	0.031
CFI	0.958	0.974	0.979	0.983
TLI	0.958	0.973	0.978	0.982

Note. RMSEA- root-mean-square error of approximation index; RMSR- root-mean-square residual; CFI- comparative fit index; TTL- Tucker-Lewis index. Values in brackets show 90% confidence interval for RMSEA.

Table 2.

EFA results based on final VFS-A-40 items.

	Fit Indices across 1–4 factor EFA			
Fit Indices	**1-Factor**	**2-Factor**	**3-Factor**	**4-Factor**
RMSEA	0.041 [0.039,0.044]	0.031 [0.029,0.034]	0.022 [0.019,0.025]	0.015 [0.011,0.018]
SRMR	0.037	0.027	0.021	0.017
CFI	0.985	0.989	0.993	0.995
TLI	0.984	0.988	0.991	0.993

Note. RMSEA- root-mean-square error of approximation index; RMSR- root-mean-square residual; CFI- comparative fit index; TTL- Tucker-Lewis index. Values in brackets show 90% confidence interval for RMSEA.

Table 3.

Mean VFS-A-40[1] summed scores and standard errors (SE) at Time 1 and Time 2 for a group of adult cochlear implant users (N=86). Wilcoxon Z and the resultant p-values in () are shown in the fourth column. Pearsons R and Intraclass correlation coefficient (ICC) values for comparisons of scores at T1 and T2 are shown in the fifth and final columns, respectively. ICC 95% confidence intervals are shown in (). **Bolded** values are significant at the 0.01 level (2-tailed).

Summed Scores	Time 1	Time 2	Wilcoxon Z	Pearsons R	ICC
Total	70.7 (3.6)	71.1 (3.5)	−.177 (.860)	**.68**	**.68 (.55–.78)**
Cognitive	22.1 (0.93)	22.6 (0.88)	−.476 (.634)	**.69**	**.69 (.57–.79)**
Emotional	17.2 (0.99)	17.0 (0.96)	−.383 (.702)	**.68**	**.69 (.56–.78)**
Physical	13.7 (0.94)	13.6 (0.93)	−.561 (.575)	**.60**	**.61 (.45–.73)**
Social	17.8 (0.99)	18.0 (0.96)	−.290 (.772)	**.68**	**.69 (.59–.78)**