Research Article

Development and Evaluation of Pediatric Versions of the Vanderbilt Fatigue Scale for Children With Hearing Loss

Benjamin W. Y. Hornsby,[a] Stephen Camarata,[a,b] Sun-Joo Cho,[b] Hilary Davis,[a] Ronan McGarrigle,[c] and Fred H. Bess[a]

[a] Department of Hearing and Speech Sciences, School of Medicine, Vanderbilt University, Nashville, TN [b] Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, TN [c] Department of Psychology, University of Bradford, United Kingdom

ARTICLE INFO

Article History:
Received January 23, 2022
Revision received March 12, 2022
Accepted March 15, 2022

Editor-in-Chief: Peggy B. Nelson
Editor: Michelle R. Molis

https://doi.org/10.1044/2022_JSLHR-22-00051

ABSTRACT

Purpose: Growing evidence suggests that fatigue associated with listening difficulties is particularly problematic for children with hearing loss (CHL). However, sensitive, reliable, and valid measures of listening-related fatigue do not exist. To address this gap, this article describes the development, psychometric evaluation, and preliminary validation of a suite of scales designed to assess listening-related fatigue in CHL: the pediatric versions of the Vanderbilt Fatigue Scale (VFS-Peds).
Method: Test development employed best practices, including operationalizing the construct of listening-related fatigue from the perspective of target respondents (i.e., children, their parents, and teachers). Test items were developed based on input from these groups. Dimensionality was evaluated using exploratory factor analyses (EFAs). Item response theory (IRT) and differential item functioning (DIF) analyses were used to identify high-quality items, which were further evaluated and refined to create the final versions of the VFS-Peds.
Results: The VFS-Peds is appropriate for use with children aged 6–17 years and consists of child self-report (VFS-C), parent proxy–report (VFS-P), and teacher proxy–report (VFS-T) scales. EFA of child self-report and teacher proxy data suggested that listening-related fatigue was unidimensional in nature. In contrast, parent data suggested a multidimensional construct, composed of mental (cognitive, social, and emotional) and physical domains. IRT analyses suggested that items were of good quality, with high information and good discriminability. DIF analyses revealed the scales provided a comparable measure of fatigue regardless of the child's gender, age, or hearing status. Test information was acceptable over a wide range of fatigue severities and all scales yielded acceptable reliability and validity.
Conclusions: This article describes the development, psychometric evaluation, and validation of the VFS-Peds. Results suggest that the VFS-Peds provide a sensitive, reliable, and valid measure of listening-related fatigue in children that may be appropriate for clinical use. Such scales could be used to identify those children most affected by listening-related fatigue, and given their apparent sensitivity, the scales may also be useful for examining the effectiveness of potential interventions targeting listening-related fatigue in children.
Supplemental Material: https://doi.org/10.23641/asha.19836154

Fatigue is characterized subjectively as a mood state and is typified by feelings of low energy, tiredness, weariness, sleepiness, or exhaustion (Hornsby et al., 2016). Subjective feelings of fatigue are generally a consequence of high levels of sustained effort applied during challenging physical and/or mental tasks. Fatigue is pervasive in daily

Correspondence to Benjamin W. Y. Hornsby: ben.hornsby@vumc.org. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

living and may occur on a frequent basis in both adults and children. Not only is fatigue prevalent among the general community, it is commonly associated with a broad array of illnesses and syndromic conditions (Hockey, 2013; Wessely et al., 1998). In fact, it is one of the most common complaints reported in primary care settings (Hockey, 2013). When high levels of effort are a result of prolonged listening activities, the resultant fatigue may be referred to as *listening-related fatigue* (Bess et al., 2020; Davis, Schlundt, Bonnet, Camarata, Bess, & Hornsby, 2021; Hornsby et al., 2021). For most healthy people, fatigue is experienced as mild and transient and can be easily resolved with no significant aftereffects. However, if the fatigue is severe, repetitive, and prolonged, as occurs in many people with chronic health conditions (e.g., multiple sclerosis, cancer, diabetes, systemic lupus, and obesity), the negative consequences can be broad and significant, impacting performance, well-being, and quality of life (Bess & Hornsby, 2014; Deluca, 2005; Hockey, 2013). For example, pediatric research suggests that children with chronic health conditions are more likely to experience fatigue-related issues such as increased school absences, poor school performance, reduced participation of daily activities, troubles with sleeping, and difficulties with interpersonal relationships (Gaba & Howard, 2002; Hockenberry-Eaton et al., 1999; Ravid et al., 2009).

There is growing evidence supporting a link between hearing loss and fatigue. This evidence suggests that the common act of listening can be a challenging experience for some adults and children with hearing loss (CHL; Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021; Holman, Hornsby, et al., 2021; Hornsby, 2013; Hornsby & Kipp, 2016; Lindburg et al., 2021). Numerous studies have reported that, compared to peers without hearing loss, adults with hearing loss frequently experience greater listening effort and listening-related fatigue across a variety of communication settings and contexts (Bess et al., 2014; Bess & Hornsby, 2014; Davis, Schlundt, Bonnet, Camarata, Bess, & Hornsby, 2021; Dwyer et al., 2019; Holman et al., 2019). Furthermore, such fatigue compromises performance in the workplace, task motivation, and quality of life (Davis, Schlundt, Bonnet, Camarata, Bess, & Hornsby, 2021; Hockey, 2013; Holman, Hornsby, et al., 2021; Hornsby et al., 2021). For example, Davis, Schlundt, Bonnet, Camarata, Bess, and Hornsby (2021) conducted focus groups in a sample of adults with varying degrees of hearing loss to better understand the extent and impact of listening-related fatigue on their daily lives. Their study revealed that listening-related fatigue was indeed a significant problem for many adults with hearing loss, especially when listening in background noise. Participants reported a wide range of fatigue experiences across multiple domains (i.e., physical, mental, emotional, and social domains) that negatively affected their quality of life.

CHL also experience listening-related fatigue and, when severe, its associated sequelae (Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021; Gustafson et al., 2018; Hornsby et al., 2014; McGarrigle et al., 2019). Compared to peers without hearing loss, CHL exhibit more stress (Bess et al., 2016), lower energy (Bess et al., 1998), greater listening effort (Hicks & Tharpe, 2002; Lewis et al., 2016; McGarrigle et al., 2019), and increased listening-related fatigue (Bess et al., 2020; Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021). Davis, Schlundt, Bonnet, Camarata, Hornsby, and Bess (2021) conducted focus groups and interviews in school-age children with moderate-to-profound hearing loss, their parents, and school professionals who work with CHL. Using qualitative data from these groups, Davis and colleagues were able to clarify and describe the fatigue experiences of CHL. Primary themes from participants were used to develop a theoretical framework for understanding listening-related fatigue in CHL. Several situational determinants of listening-related fatigue were identified, including environmental/talker factors (e.g., background noise), school-specific factors (e.g., pressure to listen/do well), and child-specific factors (e.g., motivation to listen). The theoretical framework highlighted the complex ways in which these situational determinants interact with other factors, such as the use of (and barriers to using) coping strategies, to impact listening-related fatigue in CHL.

Given that fatigue and its negative consequences are common in CHL, the need for a reliable and valid measure to assess this construct is crucial. Self-report scales are the primary metrics for assessing an individual's fatigue experience (Christodoulou, 2007). Many fatigue scales exist for adults, whereas fewer tools are available for children. Importantly, until now, none were developed to target listening-related fatigue in children. Some scales are generic (i.e., designed to assess fatigue across diverse populations). One generic measure designed for use with children is the Pediatric Quality of Life Inventory Multidimensional Fatigue Scale (PedsQL-MFS; Varni et al., 2002). Several studies have used the PedsQL-MFS to assess fatigue in CHL (Hornsby et al., 2014, 2017; Sindhar et al., 2021; Werfel & Hendricks, 2016). While these studies uniformly found that CHL experienced more fatigue than their peers without hearing loss, large between-study differences were observed, signaling the potential insensitivity of this generic tool for measuring variations in the fatigue experiences of CHL (Hornsby et al., 2014, 2017; Sindhar et al., 2021).

In contrast to generic measures, some scales are created to assess fatigue associated with specific health conditions such as cancer in adolescents (Hockenberry-Eaton et al., 1999) or hearing loss in adults (Hornsby et al., 2021). Disease-specific scales may be more sensitive than generic instruments because they assess issues that are especially relevant to those with a given health condition. As a result, they may be more useful for discerning subtle differences in the health status of individuals suffering from a given disorder and/or for detecting small, yet potentially important, changes resulting from interventions (Patrick & Deyo, 1989). For example, Holman, Drummond, and Naylor (2021) examined changes in self-reported fatigue over time in a group of adults with hearing loss who were receiving hearing aids for the first time and a control group of adults with hearing loss who did not obtain hearing aids or already had hearing aids. Fatigue ratings were obtained using two generic fatigue scales (Fatigue Assessment Scale [FAS]; Michielsen et al., 2003; Multidimensional Fatigue Symptom Inventory-Short Form [MFSI-SF]; Stein et al., 2004) and the 40-item version of the Vanderbilt Fatigue Scale for Adults (VFS-A-40; Hornsby et al., 2021). The VFS-A-40 was designed specifically to assess listening-related fatigue. Ratings were obtained at the initial assessment session and then again approximately 3, 6, and 9 months post the initial assessment. Results, as measured using the VFS-A-40, revealed a significant reduction in listening-related fatigue in the group that received their first hearing aid but not in the control group. The generic instruments (FAS and MFSI-SF) were not sensitive to changes in fatigue following the hearing aid fitting, highlighting the potential for increased sensitivity when using disease-specific scales.

Whenever possible, subjective measures should be completed by the individual who is directly experiencing the difficulty (i.e., via self-report). However, in some cases, this may not be possible or optimal, for example, when assessing infants, young children, or individuals with significant illnesses and/or cognitive or developmental impairments (e.g., Eiser & Morse, 2001; Varni et al., 2010). In these situations, useful information can be obtained from proxy respondents, such as significant others or parents. Parents and professionals who work with children may be able to provide important and complementary information about the impact of a disease or disorder on a child's experience or quality of life (Eiser & Morse, 2001; Sprangers & Aaronson, 1992; Varni et al., 2007). For example, in some cases, parents (or teachers) may be more aware of a child's fatigue-related issues than the child themselves. Consider a child with longstanding hearing loss whose listening struggles at school leave them so tired they frequently nap or rest at home after school. Given their limited experience, this may seem "normal" to the child. They may be unaware that other children may not experience these same feelings. The parent, however, can more readily compare their child's behaviors to other siblings or peers, highlighting their child's fatigue. Relevant to this study, we theorized that parents, as well as professionals who interacted regularly with CHL, could provide "value added" information regarding behaviors associated with fatigue that they have observed in the child. We hypothesized that these three sources of information (self-report and two proxy-report types) would yield important, specific, and sensitive insights regarding the listening-related fatigue experienced by children.

To summarize, many CHL are at increased risk for listening-related fatigue and its multiple cascading effects. This study was developed in response to (a) the prominence of reports of subjective fatigue among CHL, (b) the potential impact of listening-related fatigue on psychosocial well-being and life quality, and (c) the current absence of a standardized and sensitive tool for assessing listening-related fatigue in CHL. Specifically, the aim of this study was to develop and validate a suite of child-centered measures of listening-related fatigue appropriate for use in CHL—the pediatric versions of the Vanderbilt Fatigue Scale (VFS-Peds). The VFS-Peds allows for the assessment of fatigue via child self-report (VFS-C), as well as parent proxy–report (VFS-P) and teacher proxy–report (VFS-T). These scales will be critical for identifying those children at most risk for developing significant listening-related fatigue, for assessing and optimizing management strategies to reduce such fatigue, and for conducting comparative effectiveness research.

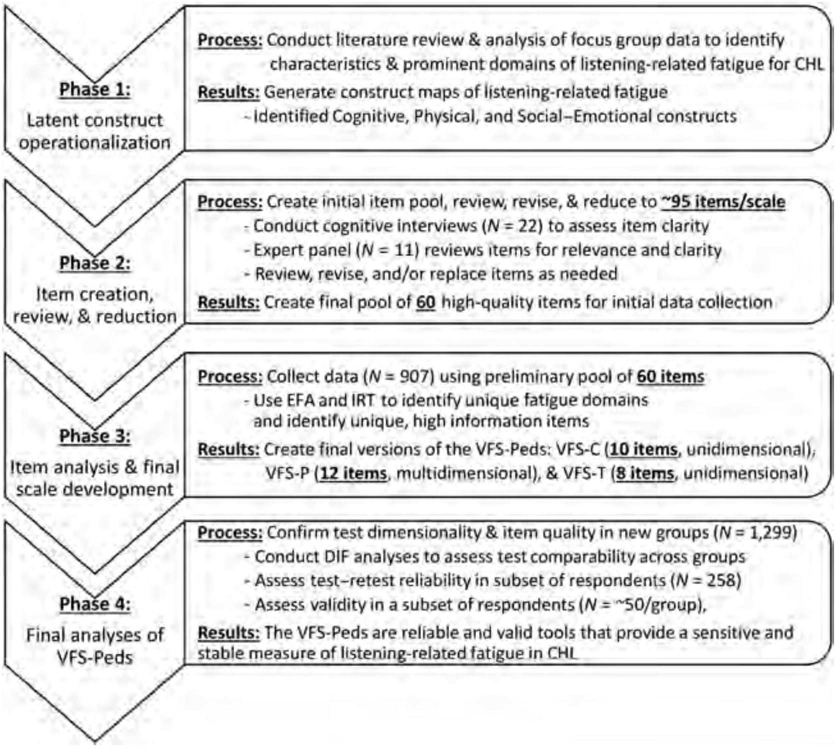
Methods

The methods used in this multiyear study were guided, in part, by our previous research developing and validating a scale to assess listening-related fatigue in adults with hearing loss (Hornsby et al., 2021). The general procedures used are often employed when developing new measures for a construct (Clark & Watson, 2019). Figure 1 provides an overview of the four phases of the development and validation process. A detailed description of these phases is provided below.

Phase 1 Methods: Latent Construct Operationalization–Listening-Related Fatigue in CHL

Phase 1 began with a review of the fatigue literature in relation to pediatric hearing loss. This review suggested that systematic research in this area was limited and comprised largely of pilot studies and anecdotal reports. To improve our understanding of fatigue in this population and to inform content for potential test items, CHL, their parents, and school professionals working with CHL were invited to participate in moderated focus groups and/or interviews to discuss the experience of listening-related fatigue in CHL. A brief description of the methods used is provided below. Details of the participants and procedures used in these focus groups have been previously published

Figure 1. Flowchart of the development and validation process for creating the pediatric versions of the VFS-Peds. CHL = children with hearing loss; EFA = exploratory factor analysis; DIF = differential item functioning; IRT = item response theory; VFS-Peds = Pediatric Vanderbilt Fatigue Scale; VFS-C = VFS child self-report scale; VFS-P = VFS parent proxy–report scale; VFS-T = VFS teacher proxy–report scale.


Phase 1:
Latent construct operationalization
Process: Conduct literature review & analysis of focus group data to identify characteristics & prominent domains of listening-related fatigue for CHL
Results: Generate construct maps of listening-related fatigue
- Identified Cognitive, Physical, and Social–Emotional constructs
Phase 2:
Item creation, review, & reduction
Process: Create initial item pool, review, revise, & reduce to ~95 items/scale
- Conduct cognitive interviews (N = 22) to assess item clarity
- Expert panel (N = 11) reviews items for relevance and clarity
- Review, revise, and/or replace items as needed
Results: Create final pool of 60 high-quality items for initial data collection
Phase 3:
Item analysis & final scale development
Process: Collect data (N = 907) using preliminary pool of 60 items
- Use EFA and IRT to identify unique fatigue domains and identify unique, high information items
Results: Create final versions of the VFS-Peds: VFS-C (10 items, unidimensional), VFS-P (12 items, multidimensional), & VFS-T (8 items, unidimensional)
Phase 4:
Final analyses of VFS-Peds
Process: Confirm test dimensionality & item quality in new groups (N = 1,299)
- Conduct DIF analyses to assess test comparability across groups
- Assess test–retest reliability in subset of respondents (N = 258)
- Assess validity in a subset of respondents (N = ~50/group),
Results: The VFS-Peds are reliable and valid tools that provide a sensitive and stable measure of listening-related fatigue in CHL

(see the work of Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021).

Participants

Participants were recruited from Vanderbilt Bill Wilkerson Center (VBWC) Audiology clinics and surrounding community. In Phase 1, a total of 88 individuals completed a focus group or interview. Participants included 43 children with bilateral, moderate-to-profound hearing loss (7–17 years; M = 12.4 years); 17 parents; and 28 school professionals working with CHL. School professionals included teachers (general educators and deaf educators), speech-language pathologists, and educational audiologists. For additional information, see the work of Davis, Schlundt, Bonnet, Camarata, Hornsby, and Bess (2021).

Procedures

Focus group discussions and interviews were recorded and transcribed for further review and analysis (approximately 21 hr of recordings were transcribed verbatim). Common themes were identified and organized to create a theoretical framework to conceptualize listening-related fatigue. See the work of Davis, Schlundt, Bonnet, Camarata, Hornsby, and Bess (2021) for a detailed discussion of these procedures. We utilized these data to develop a "construct map" (Wilson, 2005) that described the domains and severity levels of listening-related fatigue experienced by CHL from the perspectives of the child, their parents, and the school professionals working with them (see Supplemental Material S1). This construct map was used to guide the development of test items for the VFS-Peds.

Phase 2 Methods: Item Creation, Initial Item Review, Reduction, and Assessment

In Phase 2, information from focus group participants was used to create large pools of potential test items for each version of the VFS-Peds (i.e., the VFS-C, VFS-P, and VFS-T). The items were reviewed by target respondents and an expert panel. Items were revised or eliminated based on their feedback resulting in an initial bank of approximately 60 test items for each scale.

Participants

Phase 2 participants included nine CHL, six parents, and seven school professionals who completed cognitive interviews. Parents and children were again recruited from VBWC Audiology clinics, whereas school professionals were recruited via local school districts. In addition, 11 professionals who worked clinically and/or in research areas related to CHL (three speech-language pathologists, five audiologists, two deaf educators, and one hearing scientist) served as expert panel members. Each scale was reviewed for completeness, content relevance, and clarity by six (child and teacher scales) or seven (parent scale) of members of the expert panel.

Procedures

Initial item development and item reduction process. Transcripts from focus group discussions and interviews were organized for review by research laboratory members. Two to three members read each transcript, line by line, to review participant comments. If a comment provided information relevant to listening-related fatigue, the laboratory member created a potential test item, using the behavior, situation, or experience noted by the participant. This process enhances the content validity of the test items selected for the final scale (Hornsby et al., 2021). Using this approach, we created large pools of items based on comments from each respondent group (~95 items/group). These items were iteratively reviewed by the research team to remove redundant and unclear items. Relevant and clear items that remained were coded, based on the construct map, to a specific domain (physical, cognitive, and social–emotional) and severity level (mild, moderate, or severe). Our end goal was to develop a sensitive, reliable, and valid scale that was brief enough to ensure clinical utility and practicality. The number of items needed to achieve such sensitivity depends largely on the quality of the test items, but this number can be optimized using item response theory (IRT) analyses. Work on the adult version of the VFS (Hornsby et al., 2021) suggests that a sensitive test can be created with as few as approximately 10 items.

Initial item assessment via cognitive interviews and expert panel review. As a next step to identify high-quality items, we completed cognitive interviews to ensure that target respondents' (i.e., children, parents, and teachers) interpretation of potential items was consistent with our original intent. Participants were asked to describe out loud their thought process as they read and responded to each test item. This allowed the examiner to understand how the question was interpreted by the participant and to identify potential problems with interpretation or comprehension for a given item. All interviews were recorded and transcribed for later review.

The same items were also reviewed by members of our expert panel. Panel members were asked to rate each item in terms of its clarity and relevance and to provide additional feedback about items as needed. The construct map was provided for reference to help calibrate panel members understanding of the construct of listening-related fatigue. Participants rated each item for relevance using a 4-point (0–3) scale. Items were rated as *not relevant*, *somewhat relevant*, *quite relevant*, or *highly relevant*, respectively. To determine item clarity, the panel answered "yes" or "no" to the question: "Is this question well-written and easy to understand?" Findings from the cognitive interviews and expert panel review were used to reduce the initial item pool by identifying redundant and unclear items that were then removed or revised.

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

In Phase 3, we collected data from large samples of target respondents using high-quality items identified in Phase 2. These data were subjected to exploratory factor analyses (EFAs) and IRT analyses to identify high-quality items for the final versions of the VFS-Peds.

Participants

Local participants were recruited from VBWC Audiology clinics and from schools in the Middle Tennessee area. Additional participants were recruited broadly via online postings (e.g., social media postings and e-mails to groups with hearing loss). Children with, or without, any degree or configuration of hearing loss, their parents, and school professionals could participate. A total of 907 participants (see Table 1) completed the preliminary versions of the scales (60 items/scale). This consisted of 211 children aged 7–17 years (mean/median age = 13.2/13 years, *SD*/semi-interquartile range [SIQR] = 2.03/1.5 years), 392 parents, and 304 school professionals. Parents answered basic demographic information about their child, including estimating their child's binaural hearing difficulties as (a) *none*: Their child has no hearing difficulties; (b) *slight/mild*: Their child only has trouble hearing and understanding speech when it is soft, from a distance, or when listening in noise; (c) *moderate*: Their child has difficulty hearing and understanding speech at normal conversational levels, even when the speaker was relatively close; (d) *severe*: Their child, without amplification, is only able to hear loud speech or loud environmental sounds; conversational speech is generally not heard without a hearing device; or (e) *profound*: Their child generally only perceives loud sounds as vibrations. School professionals reported about their role in the education system and how many CHL were on their caseload.

Procedures

Phase 3 consisted of collecting responses to our pool of potential test items for each scale from a large sample of children with and without hearing loss, their parents, and school professionals. Data collection was completed either in person (paper forms) or online via Vanderbilt's Research Electronic Data Capture (REDCap) system

Table 1. Demographics of Phase 3 respondents.

Respondent groups	Child	Parent	School professional
Number of children reported on (*N* = 907)	*N* = 211	*N* = 392	*N* = 226/304*
Child characteristics			
Child age (mean/median & range in years)	13.2/13 (7–17)	11.2/11 (6–17)	9.8/9 (5–17)
Gender (number & percentage of sample)			
Male	98 (46%)	189 (48%)	159 (52%)
Female	105 (50%)	191 (49%)	142 (47%)
Did not disclose	8 (4%)	12 (3%)	3 (1%)
Group type (number & percentage of sample)			
Hearing loss	156 (74%)	296 (76%)	220 (72%)
Unilateral loss	34 (22%)	60 (20%)	DNC
Bilateral loss	122 (78%)	236 (80%)	DNC
No hearing loss	50 (24%)	93 (24%)	61 (20%)
Did not disclose	5 (2%)	3 (< 1%)	23 (8%)
Description of child's hearing impairment**	*n* = 156	*n* = 296	
Slight/mild	34 (22%)	65 (22%)	DNC
Moderate	60 (38%)	115 (39%)	DNC
Severe	23 (15%)	50 (17%)	DNC
Profound	31 (20%)	59 (20%)	DNC
Did not disclose	8 (5%)	7 (2%)	DNC
Parent characteristics			
Highest level of education			
Some high school	NA	12 (3%)	DNC
High school grad/GED	NA	47 (12%)	DNC
Some college/postsecondary	NA	61 (16%)	DNC
Associate's degree	NA	29 (7%)	DNC
Bachelor's degree	NA	97 (25%)	DNC
Graduate/professional degree	NA	103 (26%)	DNC
Did not disclose	NA	43 (11%)	DNC
School professional characteristics			
Professional type			
General educator	NA	NA	72 (32%)
Deaf educator	NA	NA	79 (35%)
Special educator	NA	NA	19 (8%)
Speech-language pathologist	NA	NA	23 (10%)
Educational audiologist	NA	NA	12 (6%)
Other***	NA	NA	21 (9%)

Note. DNC = did not collect data from respondents; NA = category is not applicable to respondents; GED = General Educational Development.

*Forty-five of the 226 participants provided data on multiple children. Forty-four of the 45 respondents reported on two to six children (mean/median = 2.8/2 children). One respondent reported on 18 children. **See text for descriptions. Respondents rated hearing difficulties when listening bilaterally (not ear specific). ***Occupational therapists, educational interpreters, counselors, and paraprofessionals.

(Harris et al., 2009). REDCap is a secure, web-based software platform designed to support data capture for research studies. For online data collection, parents received an electronic link to access consent and assent forms and survey items, which they completed. Parents were allowed to assist their child while completing the forms if needed. In-person data collection was completed in conjunction with a scheduled audiology clinic visit. A researcher was present and available to answer questions and observe the child as they responded to the scale. If the child requested help and/or the researcher felt the child required support, the scale questions and answer choices were read to the child. EFA was used to examine the latent structure of each scale's item pool. To identify the most high-quality items, IRT analyses were used to examine the characteristics of potential test items. A description of the analysis methods is provided below.

EFA. We first investigated the latent structure (i.e., the number of dimensions and factor loading patterns) of the child, parent, and teacher items by conducting a series of EFAs, one for each item group, using Mplus Version 8.3 (Muthen & Muthen, 1998–2017). Using polychoric correlations for ordered polytomous responses (specifically, weighted least square with adjusted means and variance with Oblimin rotation and Oblique type), a series of EFAs were conducted, extracting one to three factors. Fit indices were compared across one- to three-factor models. Guidelines used to assess the goodness of model fit included 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 (TLI; Tucker & Lewis, 1973) > .95 (Hu & Bentler, 1999; Yu, 2002).

IRT analyses. We evaluated individual item quality, as well as the quality of sets of items, using IRT analyses. Item and test quality refers to the ability of the items, or test, to precisely measure the construct of interest, which is, in this case, listening-related fatigue. Depending 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.[1] Item parameter estimates were obtained using a (marginal) maximum-likelihood estimation (MLE) method in Mplus Version 8.3 (Muthen & Muthen, 1998–2017). For IRT scoring, expected a posteriori (EAP) scoring for a specific response pattern was used. "Missingness" in item responses was considered missing at random and treated as missing under the MLE.

Three criteria were used to evaluate item quality. (a) Item thresholds needed to be ordered and well-separated (e.g., as the level of the latent construct increased, respondents selected higher response options). (b) Item discrimination needed to be positive and high in magnitude, suggesting that the item was sensitive to variations in listening-related fatigue. (c) The average item threshold needed to match their hypothesized severity level (i.e., mild, moderate, or severe) based on the construct map. These item characteristics are important as they will impact the sensitivity of the final scale.

To assess item quality, we looked at each item's information function. These functions describe the amount of information an item provides across different levels of the latent construct (denoted by θ, i.e., listening-related fatigue in our case). Our goal was to identify items with high information over a wide range of fatigue levels. We also conducted differential item functioning (DIF) analyses on all test items. DIF analyses determine whether a test item measures listening-related fatigue equivalently across distinct groups (e.g., across age and gender groups). This information was used to exclude DIF items from the final versions of the scales. 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 DIF effect is constant across trait levels) and nonuniform effect (the DIF effect varies across trait levels) and a total DIF effect.[2] 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).

Finally, we evaluated test information for various sets of high-quality items to select items for the final versions of the VFS-Peds. 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 test information for several sets of items by examining their test information curves (TICs). TICs show test information as a function of θ. High test information and a broad TIC implies that the test has good measurement fidelity across a wide range of θ's. Our goal was to identify sets of items (one set for each version of the scale) that provided a test information level of at least 11.11 (test information = 1 / [standard error of an IRT scale score^2]) over a wide range. A standard error of 0.3 was used as an empirical cutoff to calculate the target test information value of 11.11. This error value corresponds to a reliability coefficient of .95, which has been deemed acceptable in the development of other clinical scales (e.g., Cole et al., 2012; Hospers et al., 2016). In addition, we selected high information items so that we could obtain this sensitivity with a relatively small number of test items (e.g., ~10 items/scale) to enhance the clinical utility of our measures. Using this approach, we identified a reduced set of high-quality items for inclusion in the final versions of the VFS-Peds.

[1] The item discrimination parameter is a measure of the item's ability to discriminate between various levels of the construct (e.g., varying levels of fatigue severity). A larger value indicates more sensitivity to variations in the latent construct. The item threshold parameters for each item are used to model response scores. For our data set having five response scores, four-item threshold parameters are estimated. Threshold 1 reflects the transition point from a score of 0 (i.e., *never*) to scores 1–4 (*rarely*, *sometimes*, *often*, or *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. Threshold 4 reflects the transition point from scores of 0–3 to a score of 4.

[2] 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 *nonuniform 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, alpha = 0.01, was chosen.

Phase 4 Methods: Field Testing and Validation of the VFS-Peds

In Phase 4, we replicated EFA and DIF analyses and examined item and test information of the final versions of the VFS-Peds in a new sample of participants. We also examined the test reliability and validity of these final versions.

Participants

In total, 1,299 participants completed Phase 4 including 225 children, 532 parents, and 542 school professionals. Participant recruitment methods were the same as those utilized in Phase 3. Child participants ranged in age from 6 to 17 years (mean/median = 13.2/14 years; *SD*/SIQR = 2.36/2.0 years). Within the Phase 4 sample, a subset of participants completed their respective versions of the scales at two separate time points to assess test–retest reliability (*N* = 256 participants, *n* = 43 children, *n* = 139 parents, and *n* = 74 school professionals). In addition, to assess construct validity a subset of the Phase 4 respondents completed (a) a generic fatigue measure (the PedsQL-MFS) and (b) a measure of depression (the Children's Depression Inventory [CDI-2]; Kovacs & Multi-Health Systems Staff, 2011) as part of the validation process of the final versions of the VFS-Peds. Approximately, 50 CHL and their parents, as well as approximately 50 school professionals, completed these measures (*n* = 44–52 respondents for the various measures). Additional demographic/descriptive information for Phase 4 respondents is shown in Table 2.

Procedures

The EFA and IRT analyses (using Mplus) described in Phase 3 were repeated using these new Phase 4 data. In addition to item parameter estimates, and item and test

Table 2. Demographics of Phase 4 respondents who completed the final versions of VFS.

Respondent groups	Child	Parent	School professional
Number of children reported on (*N* = 1,299)	*N* = 225	*N* = 532	*N* = 532/542*
Child characteristics			
Child age (mean/median & range in years)	13.2/14 (6–17)	10.9/11 (6–17)	10.3/10 (6–17)
Gender (number & percentage of sample)			
Male	120 (53%)	278 (52%)	282 (52%)
Female	105 (47%)	253 (48%)	250 (47%)
Did not disclose	0 (0%)	1 (< 1%)	10 (1%)
Group type (number & percentage of sample)			
Hearing loss	119 (53%)	305 (57%)	368 (68%)
Unilateral loss	35 (29%)	77 (25%)	55 (15%)
Bilateral loss	84 (71%)	228 (75%)	313 (85%)
No hearing loss	106 (47%)	227 (43%)	170 (32%)
Did not disclose	0 (0%)	0 (0%)	4 (< 1%)
Description of child's hearing impairment**	*n* = 119	*n* = 305	
Slight/mild	23 (19%)	59 (19%)	DNC
Moderate	34 (29%)	85 (29%)	DNC
Severe	25 (21%)	56 (18%)	DNC
Profound	35 (30%)	100 (33%)	DNC
Did not disclose	2 (1%)	5 (1%)	DNC
Parent characteristics			
Highest level of education			
Some high school	NA	17 (3%)	DNC
High school grad/GED	NA	38 (7%)	DNC
Some college/postsecondary	NA	79 (15%)	DNC
Associate's degree	NA	42 (8%)	DNC
Bachelor's degree	NA	149 (28%)	DNC
Graduate/professional degree	NA	203 (38%)	DNC
Did not disclose	NA	4 (1%)	DNC
School professional characteristics			
Professional type			
General educator	NA	NA	143 (26%)
Deaf educator	NA	NA	213 (39%)
Special educator	NA	NA	53 (10%)
Speech-language pathologist	NA	NA	63 (12%)
Educational audiologist	NA	NA	27 (5%)
Other***	NA	NA	43 (8%)

Note. VFS = Vanderbilt Fatigue Scale; DNC = did not collect data from respondents; NA = category is not applicable to respondents; GED = General Educational Development.

*Ten school professional respondents provided data on two students (one child with and one without hearing loss). **For bilateral losses, respondents rated the level of hearing difficulty in the better hearing ear. ***Therapists, counselors, paraprofessionals, and those who indicated more than one profession.

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 DIF analyses across gender, age, grade level, and self-reported hearing loss groups. Test–retest reliability was assessed in a subset of the original participant pool. Phase 4 respondents interested in participating in the retest session provided their e-mail address, and a link to the retest was e-mailed to them approximately 1–2 weeks after they completed the initial scale.

Finally, we assessed construct validity in a subset of Phase 4 respondents by examining associations between their VFS-Peds scores and scores obtained using the PedsQL-MFS and the CDI-2. Given that the VFS-Peds and the PedsQL-MFS both assess the construct of fatigue, as evidence of convergent validity, we expected moderate correlations between these measures. As evidence of discriminant validity, we expected somewhat weaker correlations with the CDI-2 given that, while frequently co-occurring, fatigue and depression are often considered distinct constructs (Johnson, 2005; Michielsen et al., 2003; Wilson et al., 2018). The PedsQL-MFS consists of child self-report and parent proxy–report scales. There is no teacher version of the PedsQL-MFS; hence, no teacher data were analyzed. Each version assesses three fatigue domains—general, cognitive, and sleep/rest fatigue. The various versions of the PedsQL-MFS have been shown to have good internal consistency and construct validity (Varni et al., 2002, 2007). Depression was assessed by the child using the child self-report screening version of the CDI-2, whereas parents and teachers/school professionals completed the parent and teacher versions, respectively. All versions of the CDI-2 provide a total score. The parent and teacher versions also provide subscale scores reflecting emotional and functional problems. Raw scores are converted to *T*-scores ($M = 50$, $SD = 10$) for interpretation. The versions of the CDI-2 have high or acceptable internal consistency and good short-term test–retest reliability (Kovacs & Multi-Health Systems Staff, 2011). When examining associations, Bonferroni corrections were used to control for multiple comparisons. The number of comparisons varied across scales (see Table 6) such that adjusted significance levels ranged from .0167 (for VFS-T associations; $p = .05/3$) to .0071 (for VFS-P associations; $p = .05/7$). The strength of the correlations was rated as *weak* (< 0.4), *moderate* (0.4 to 0.7), or *strong* (≥ 0.7; Akoglu, 2018).

Results

Phase 1 Results: Latent Construct Operationalization

Our analysis of focus group transcripts revealed that the construct of listening-related fatigue in CHL shared many similarities to that described by adults with hearing loss. However, important differences were also observed (see the work of Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021, for discussion). Similar to adults, the experience of listening-related fatigue in CHL was complex with multiple expressions in the physical, cognitive, and social–emotional domains. Differing somewhat from the adult population, the experience of fatigue in the social and emotional domains was more closely linked in CHL, resulting in a combined social–emotional domain. Also, similar to results obtained from adults, the severity of fatigue across domains varied widely between children. Some CHL reported no fatigue-related issues or only mild levels of fatigue in isolated domains, whereas others reported severe problems across all domains (see the work of Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021, for details). We used this information to create a construct map that operationalized listening-related fatigue in CHL based on the feelings and adaptive behaviors reported by the children, parents, and/or school professionals (see Supplemental Material S1). It should be noted that despite the large situational differences between adults with hearing loss and CHL (e.g., school vs. work settings), their construct maps overlapped considerably. In the next phase, guided by the construct map, information from focus groups was used to develop relevant, potential VFS-Peds test items.

Phase 2 Results: Item Creation, Initial Item Review, Reduction, and Assessment

Initial item development and item reduction process. Transcripts from Phase 1 focus groups and interviews provided over 2,800 quotes from children, parents, and school professionals that were used to guide the construction of a large pool of potential test items. A subset of the study authors ($N = 3$) met to review these initial items and to exclude/revise poorly worded or redundant items. This process resulted in an initial pool of 586 items (212, 182, and 192 potential test items for the child, parent, and teacher scales, respectively). These potential test items were subjected to further review by the full research team. Specifically, each team member evaluated all 586 items and, individually, selected approximately 300 high-quality items (~95 items for each scale). Team members were instructed to select items (a) to ensure coverage of all domains/severity levels as described in the construct map and (b) based on the item's readability, clarity, and uniqueness. During this process, team members also edited items in order to ensure simple syntax and clarity, particularly for the child scale. The research team then met as a group to review the individual selections and reach consensus on a further reduced pool of items. This iterative review process resulted in a reduced item pool containing approximately 95 high-quality items (91, 94, and 96 items for the child, parent, and teacher scales, respectively). Response options for all test items used a 5-point Likert frequency scale including options of *never*, *rarely*, *sometimes*, *often*, and *almost always*. This pool of potential test items was then subjected to further evaluation via cognitive interviews and expert panel review.

Assessment via cognitive interviews and expert panel review. Results from cognitive interviews revealed that most items were clearly understood and interviewee responses to items were consistent with the item's underlying intent. However, based on recommendations from interviewees, several items were modified to improve clarity. These items were then provided to an expert panel for further evaluation. Expert panel members rated the relevance and clarity of all potential test items for each scale and the comprehensiveness of the item pools in relation to the construct map. Items were revised as needed to address concerns regarding clarity or relevance raised by panel members. In cases where panel members provided conflicting information, the research team met and reviewed the concerns to reach consensus on modifications or to delete the test item. Our goal was to use the information from cognitive interviews and the expert panel review to identify a smaller pool of the most relevant and clear items that would be used for data collection in Phase 3.

Based on comments from expert panel members and further review by team members, the item pool for each scale was reduced from approximately 95 items per scale to 60 items per scale. Each version of the VFS (child, parent, and teacher) contained 20 items in each of three domains: physical, cognitive, and social–emotional. Items specific to hearing device users were removed from the pool of potential test items based on our experiences with the adult scale. Inclusion of device-specific items creates scoring issues when the test is used with users and nonusers of hearing devices (Hornsby et al., 2021). To ensure that the items selected for the final scales were sensitive to differences in moderate-to-severe listening-related fatigue, more items were chosen within each domain to target moderate (seven items) and severe (eight items) listening-related fatigue than mild fatigue (five items). This approach was used for all three versions of the VFS. In Phase 3, data were collected and analyzed using these 60-item versions of the VFS.

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

Initial EFA. The purpose of these initial analyses was twofold. First, we wanted to explore the underlying factor structure of listening-related fatigue for each respondent group. Our second goal was to evaluate the quality of individual items and use this information to guide the selection of a smaller number of high-quality items for use in a final, clinically feasible, version of the VFS-Peds.

Results of the initial EFA suggested that the factor structure varied across respondent groups (see Table 3). We evaluated goodness of fit for the one-, two- and three-factor models according to the model–data fit indices described above. Consistent with a unidimensional model, almost all test items for the child self-report data loaded highly onto a single factor and goodness-of-fit indices confirmed that a one-factor model fit the child data well. For

Table 3. Phase 3 EFA results based on 60-item pools.

	Fit index		
Results	**1-factor**	**2-factor**	**3-factor**
Child self-report			
RMSEA	0.054 [0.051, 0.058]	0.045 [0.041, 0.049]	0.039 [0.034, 0.043]
RMSR	0.066	0.053	0.046
CFI	0.953	0.969	0.978
TLI	0.951	0.967	0.975
Teacher proxy–report			
RMSEA	0.091 [0.089, 0.094]	0.077 [0.075, 0.080]	0.065 [0.063, 0.068]
RMSR	0.080	0.059	0.042
CFI	0.938	0.957	0.971
TLI	0.936	0.954	0.967
Parent proxy–report			
RMSEA	0.090 [0.088, 0.092]	0.075 [0.073, 0.077]	0.064 [0.061, 0.066]
RMSR	0.078	0.063	0.049
CFI	0.921	0.947	0.963
TLI	0.918	0.943	0.959

Note. Values in brackets show 90% confidence interval for RMSEA. EFA = exploratory factor analysis; RMSEA = root-mean-square error of approximation index; RMSR = root-mean-square residual; CFI = comparative fit index; TLI = Tucker–Lewis index.

the teacher data, a one-factor model fit the data reasonably well, although the fit was not as robust as that seen in the child data. RMSEA values for the teacher data did not meet our criterion cutoff of < 0.06 regardless of the model structure. However, the other indices suggested that a one-factor model provided a reasonable fit and was more compelling than a two- or three-factor solution. Specifically, while not exceeding our a priori goodness of fit criteria, RMSR, CFI, and TLI indices all reasonably approximated these cutoffs. In addition, while a two-factor model provided a slightly better fit to our data (see Table 3), a review of individual item factor loadings for a one-factor EFA revealed that all 60 test items loaded significantly on the single factor. An analysis of item factor loadings for the two-factor EFA revealed many items (50 of the 60 items) loaded significantly on both factors; hence, an orthogonal set of items for each factor could not be constructed. Only four items loaded heavily on the second factor alone and the wording of three of these items were quite similar (all included the phrase that the student was "more tired" than their peers and thus could be viewed as overlapping versions of a single "unique" item). Given that goodness-of-fit indices suggested that a one-factor model provided a reasonable fit and the limited number of test items that loaded uniquely on a second factor, we felt that a one-factor model provided an acceptable fit to our teacher data.

In contrast, the EFA of parent data clearly suggested that a multidimensional model provided the most appropriate fit. The analyses also revealed relatively distinct items for each factor. Most cognitive, social, and emotional items loaded onto one omnibus "mental" factor, whereas the physical items loaded primarily onto a second "physical" factor, thus providing a clear two-factor solution. Moreover, with the exception of RMSEA values, goodness-of-fit indices suggested that a two-factor model fit the parent proxy data reasonably well with all indices closely approximating cutoff criteria. Similar to the teacher data, goodness of fit improved by adding another factor (i.e., a three-factor model). However, once again, a review of individual item factor loadings did not reveal a clear unique three-factor structure to our item pool. For example, all items that loaded significantly on the third factor also loaded significantly on at least one other factor. Only two items loaded more significantly on the third factor than on either the first or the second factor, essentially limiting the number of items available for assessing a third factor to two. Together, these findings suggested that a two-factor model provided an acceptable fit to our data while also providing enough items loading onto each factor to develop a reliable parent proxy–report scale.

Initial item assessment and reduction using IRT analyses. Based on the EFA results, a unidimensional GRM was fit to the child self-report and teacher proxy–report data sets, whereas a multidimensional GRM was fit to the parent data. We first conducted IRT analyses to examine the item parameter estimates of the GRM. The item parameter estimates are on the logit scale. Evaluating our 60-item pools, across all scales, item discrimination estimates ranged from −0.06 to 5.5 (M = 2.35). We also examined each item's threshold separation and threshold ordering. High-quality items are those with high item discrimination and whose thresholds are in order and well separated. Most of the items analyzed in this phase had moderate-to-high discrimination values (i.e., 97% of items had discrimination values > 1.0) and were well-ordered with good threshold separation.

Next, DIF analyses were conducted to determine whether potential test items measured listening-related fatigue equivalently across distinct subpopulations. For this analysis, it is important that respondents utilize the full range of response options available for a given item (i.e., respondents select ratings ranging from *never* to *almost always*). Thus, test items that did not have at least five responses in each given category were excluded from further evaluation. This criterion resulted in the exclusion of 16 items (six child, 10 parent, and no teacher items) from further evaluation. We then assessed DIF of the remaining test items, for each respondent group, as a function of gender (self-reported male or female) and age. For the age comparison, we used the median age of the children reported on in each group as the cutpoint for comparisons. Briefly, items were flagged as a DIF item when any of the three likelihood ratio χ^2 statistics were significant (see Phase 3 Methods section). Results revealed a total of 28 DIF items. There were 14 DIF items for age alone, 11 DIF items for gender alone, and three items for both age and gender. Thus, a total of 44 items were excluded from the original 180 potential test items (60/respondent group) based on DIF (28 items) and/or inadequate use of the range of response option (16 items). This left a total of 51, 43, and 42 potential test items for the child, parent, and teacher scales, respectively.

Selection of final test items. Our goal was to create a sensitive and reliable scale with a relatively small number of test items to enhance its clinical utility. An iterative process incorporating IRT analysis methods (i.e., item and TICs), prior comments from our expert panel review (as described earlier), and ongoing input from the research team was used to reduce the item pools and select final test items. Our goal for test sensitivity was a test information level of at least 11.1 over a wide range of severity levels (θ's). We evaluated various sets of items and their resulting TICs to identify item sets for the final scales.

The process described above resulted in (a) a single-factor, 10-item self-report scale for the child (VFS-C); (b) a single-factor, eight-item teacher proxy–report scale (VFS-T); and (c) a two-factor (mental and physical), 12-item parent proxy–report scale (VFS-P). The Phase 3 EFAs on the child and teacher test items suggested a unidimensional structure for these scales. Therefore, test items for the VFS-C and VFS-T were chosen based on their item properties and feedback from cognitive interviews and expert panel reviews. No attempt was made to include items based on their assumed fatigue domains (i.e., physical, cognitive, or social–emotional). In contrast, EFA results for the parent items suggested a two-factor structure. Thus, for the VFS-P, in addition to using IRT and qualitative results to select items, we also selected items that were highly loaded onto one of the two factors. As a result, the final version of the VFS-P contained seven items that loaded highly on a "mental" factor (VFS-P$_{\text{Mental}}$; items assessing cognitive and social–emotional aspects of listening-related fatigue) and five items that loaded on a "physical" factor (VFS-P$_{\text{Physical}}$; items assessing sleep/rest issues and/or general tiredness). In Phase 4, the sensitivity and validity of these final scales were evaluated using data collected from 1,299 parent, child, and teacher respondents.

Table 4. Pediatric versions of the Vanderbilt Fatigue Scale (VFS-Peds) test items, including the child self-report (VFS-C), parent proxy–report (VFS-P), and teacher proxy–report (VFS-T) scales, respectively.

VFS-C

1. I want to "zone out" in very noisy places.
2. It is hard for me to concentrate when lots of people are talking.
3. My brain gets tired after listening all day.
4. I get worn out from listening at school.
5. Trying to listen at school stresses me out.
6. I use a lot of energy trying to listen in class.
7. I want to go to sleep after a long day of listening.
8. I give up trying to listen when I get tired.
9. I get so tired from listening that I don't want to do anything else.
10. I feel worn out when I have to listen carefully.

VFS-P

Mental fatigue

1. My child gets frustrated when it is difficult to hear.
2. My child prefers to be alone after listening for a long time.
3. My child shuts down after listening for a long time.
4. Listening takes a lot of effort for my child.
5. My child gets tired of listening by the end of the day.
6. My child shuts down if he/she becomes frustrated from listening.
7. My child "gives up" in difficult listening situations.

Physical fatigue

8. My child needs time to relax after school.
9. My child is so tired that he/she lays down to rest.
10. My child seems drained at the end of the school day.
11. My child is more tired during the week than on weekends.
12. My child needs to relax after a tiring day of listening.

VFS-T

1. The student will "check out" after long periods of listening.
2. The student seems less motivated to do work after listening for a long time.
3. The student stops participating when struggling to hear.
4. The student seems to get worn out from listening all day at school.
5. The student has trouble concentrating when it is difficult to hear.
6. The student seems to give up more easily when having trouble listening.
7. The student appears irritated when it is hard to hear and understand.
8. The student needs listening breaks in order to stay on task.

Phase 4 Results: Analysis and Validation of the Final Versions of the Pediatric VFS-Peds

The final versions of the scales comprising the VFS-Peds are provided in Supplemental Materials S2–S4. The test items for each scale are shown in Table 4. The mean Flesch–Kincaid grade reading levels and range of individual items (shown in parentheses) for VFS-Peds were 3.6 (1.2–5), 3.9 (0.5–8.1), and 6.9 (4–8.1) for the child, parent, and teacher scales, respectively. All scales use 5-point, Likert frequency response options that range from *never* to *almost always*. Scoring can be completed by simply summing the numerical values (0–4) for the five Likert ratings. This scoring process results in a single "total" listening-related fatigue score for the unidimensional VFS-C and VFS-T and two subscale scores for the VFS-P (i.e., VFS-P$_{\text{Mental}}$ and VFS-P$_{\text{Physical}}$).

In addition, all scales also allow for IRT scoring, which takes the sensitivity (i.e., item discrimination) of items into account when calculating a scale score. This method provides a more precise estimate of an individual's listening-related fatigue (see Hambleton & Swaminathan, 2013, for review). When using IRT scoring, the resultant VFS scores are a type of scale score, similar to *z* scores, but on a logit scale. A VFS IRT scale score of 0 reflects a score that is equal to the mean IRT score of the sample. An individual VFS IRT scale score that is highly negative or positive (e.g., −3 or +3) would suggest that the individual had very low or very high listening-related fatigue, respectively.

An analysis of missing item responses suggests that the scales were easy for participants to use and complete. When looking at individual items across all scales, < 1% of items had missing responses (range of 0%–0.94%). In addition, across individuals, the vast majority of participants provided a response to all test items (98.2%, 95.5%, and 98.2% for the child, parent, and teacher scales, respectively). The maximum number of items missed by any individual was three, and this occurred only once (a teacher respondent). In the rare cases where respondents did not complete an item, in most cases (95.9%–98.7%), only one item was missed.

Reassessment of Factor Structure and Item and Test Characteristics

EFA. Using data from Phase 4 respondents, we repeated our EFA analyses. EFA model fit indices (except for RMSEA estimates) on this new data set, obtained using the final versions of the VFS-Peds, continued to suggest a one-factor model fit for the teacher proxy–report and child self-report data and a two-factor solution for the parent proxy–report data. Model fits based on RMSEA estimates were suboptimal across all scales (see Table 5). In contrast, all other indices and patterns in item factor loadings suggested that our data were modeled well using a one-factor (child & teacher) or a two-factor (parent) solution. Consistent with the initial EFA, VFS-P items assessing cognitive, social, and emotional aspects of listening-related fatigue (n = 7) loaded highly onto one "mental" factor (VFS-P_{Mental}), whereas items assessing "physical" aspects of listening-related fatigue (n = 5) loaded more highly onto a separate factor (VFS-P_{Physical}).

IRT analyses. Based on these EFA results, we again used a unidimensional GRM to fit the child self-report and teacher proxy–report data and a multidimensional GRM to fit the parent proxy–report data. Results revealed that item discrimination estimates remained high, with mean values of 1.95, 2.91, and 2.94 for the child, teacher, and parent scales, respectively. The range of item discrimination scores, across scales, varied from 1.14 to 4.03. Thresholds for all items were well ordered and adequately separated. Figure 2 shows examples of category characteristic curves for three representative items, one from each version of the VFS-Peds.

Item parameter (discrimination and threshold) estimates of the GRM (and item responses) can be used to calculate individual IRT scale scores based on all items (for the VFS-C and VFS-T) and IRT subscale scores (for the VFS-P_{Mental} and VFS-P_{Physical}). The item parameter estimates for the child, parent, and teacher scales are provided in Supplemental Material S5. Based on the results of generalized χ^2 tests, all items were fit well by the GRM. We also examined test information using data from Phase 4 respondents. Figure 3 shows TICs for each of the final VFS-Peds scales. These TICs show that the VFS-Peds have good fidelity (test information ≥ 11.11) over a wide range of severities of listening-related fatigue (i.e., IRT scale scores). Averaged across all scales, we see good fidelity for θ's ranging from −1.6 to 1.5. This range, however, varies somewhat across scales. For example, the TIC for the physical factor of the parent scale (VFS-P_{Physical}) shows good fidelity for people with θ's ranging between approximately −1.6 and 1.1, the narrowest range of any of the VFS-Peds. In contrast, the VFS-T provides good fidelity over a wider range of θ's (−2.2 to 1.7). For context, over 90% the IRT scale scores of Phase 4 participants using the child (VFS-C; 93.8%), teacher (VFS-T; 90.2%), and "mental" components of the parent scales (VFS-P_{Mental}; 94.7%) were below the upper θ boundaries (i.e., where test information fell below 11.1). For the VFS-P physical subscale (VFS-P_{Physical}), a smaller proportion of respondents (68.4%) fell below the upper fit boundary.

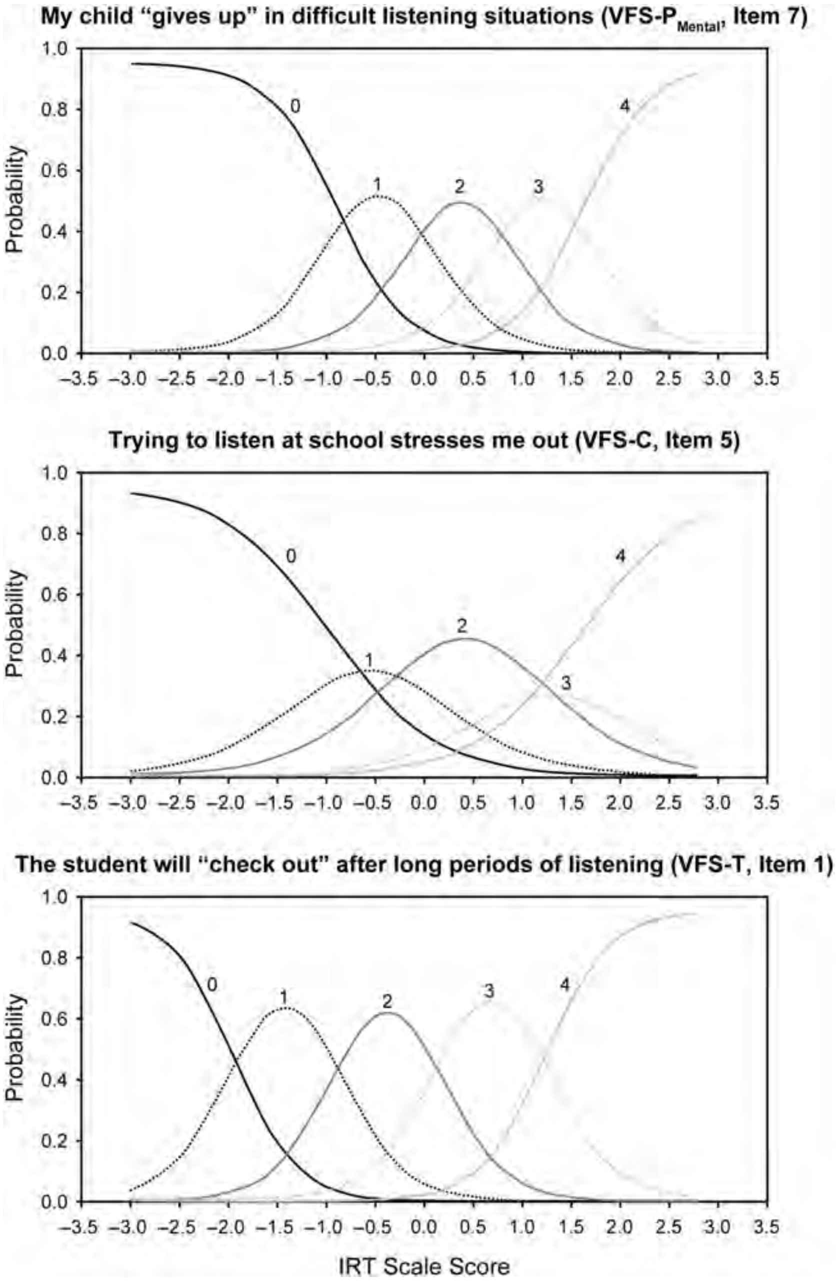
Finally, we replicated our DIF analyses, as described in Phase 3, examining gender, age, grade level, and self-reported hearing loss. When examining DIF across ages, the median age of the children from this sample was used as a cutpoint (median age = 11, 14, and 10 years for the parent, child, and teacher groups, respectively). When examining DIF across grade levels, we chose cutpoints based on (a) "natural" grade breaks and (b) an attempt to have similar numbers of respondents in each group. This resulted in different cutpoints for our parent and teacher analyses and the child analyses. A cutpoint of 4th grade (kindergarten through 4th grade vs. 5th grade through high school) was used for the parent and

Table 5. Phase 4 EFA results based on final versions of the VFS.

	Fix index		
Results	**1-factor**	**2-factor**	**3-factor**
Child self-report			
RMSEA	0.114 [0.094, 0.134]	0.110 [0.087, 0.134]	0.110 [0.083, 0.139]
RMSR	0.064	0.049	0.035
CFI	0.964	0.975	0.983
TLI	0.954	0.957	0.956
Teacher proxy–report			
RMSEA	0.139 [0.122, 0.157]	0.064 [0.041, 0.088]	0.044 [0.000, 0.079]
RMSR	0.041	0.016	0.999
CFI	0.983	0.998	0.998
TLI	0.976	0.995	0.010
Parent proxy–report			
RMSEA	0.184 [0.174, 0.194]	0.098 [0.087, 0.109]	0.077 [0.064, 0.091]
RMSR	0.105	0.032	0.021
CFI	0.939	0.986	0.993
TLI	0.926	0.979	0.987

Note. EFA = exploratory factor analysis; VFS = Vanderbilt Fatigue Scale; RMSEA = root-mean-square error of approximation index; RMSR = root-mean-square residual; CFI = comparative fit index; TLI = Tucker–Lewis index.

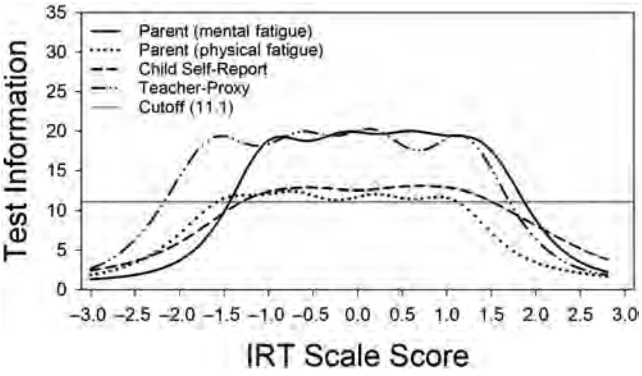
Figure 2. Exemplar category characteristic curves for representative items from the VFS-Peds. Curves labeled 0, 1, 2, 3, and 4 represent the probability a participant (i.e., a child, parent, or teacher) responds using the option *never*, *rarely*, *sometimes*, *often*, or *almost always*, respectively, given the child's underlying magnitude of listening-related fatigue as indicated by the IRT scale score. IRT = item response theory; VFS-C = VFS child self-report scale; VFS-P = VFS parent proxy–report scale; VFS-Peds = pediatric versions of the Vanderbilt Fatigue Scale; VFS-T = VFS teacher proxy–report scale.



teacher data, whereas a cutpoint 8th grade (kindergarten through 8th grade vs. 9th grade through high school) was used for analysis of child data. When examining DIF based on self-reported hearing loss, participants were grouped based on the parents' or teachers' response to the yes-or-no question, "Does your child [or, for teachers, "your student"] have a hearing loss?"

Items were flagged as DIF items when any of the three (uniform, nonuniform, or total) likelihood ratio χ^2 statistics were significant. Across all scales and based on one of the three likelihood ratio χ^2 statistics, there were two DIF items for gender, four DIF items for age, five DIF items for grade level, and six DIF items for hearing loss. However, the DIF effect sizes (McFadden's pseudo

Figure 3. Test information curves for the VFS-Peds, including the VFS-P (mental and physical domains), the VFS-C, and the VFS-T. Dashed horizontal line represents the minimum acceptable test information level (i.e., test information = 11.1). IRT = item response theory; VFS-C = VFS child self-report scale; VFS-P = VFS parent proxy–report scale; VFS-Peds = pediatric versions of the Vanderbilt Fatigue Scale; VFS-T = VFS teacher proxy–report scale.



R^2 measure) for all these items were negligible (average and maximum effect sizes of 0.015 and 0.038, respectively). These results suggest that VFS-Peds scores can be interpreted the same way in younger and older children, children in elementary school and high school, in males and females, and in children with and without hearing loss. These findings, in conjunction with our rigorous item development and evaluation process, provide good evidence of VFS-Peds test validity.

Construct validity. We further assessed validity by examining correlations between VFS-Peds IRT scale scores, the PedsQL-MFS, and the CDI-2. Associations between each version of the VFS-Peds and the various PedsQL-MFS and CDI scores were examined separately. There is no PedsQL-MFS scale for teachers; therefore, no comparisons with teacher data were made. Likewise, for the CDI child self-report scale, the only association examined was between VFS-C and a total CDI-2 score as the instrument does not provide subscale scores. In addition, CDI-2 *T*-scores for five teachers were categorized as "≤ 40" as their raw scores were 0. For these participants, a *T*-score of 40 was used in the correlation analyses. We replicated the analyses excluding these five participants, and the pattern of significant results did not change. Results of these analyses are shown in Table 6.

We anticipated moderate, negative correlations between VFS-Peds and PedsQL-MFS scores given that both are measures of the broad construct of fatigue. Consistent with this expectation, almost all correlations between VFS-P (mental and physical factors) and PedsQL-MFS scores were moderate, or stronger, in magnitude. The only exception was a weak, but still significant, negative correlation between VFS-P_{Mental} and PedsQL sleep/rest scores. Correlations between the VFS-C and PedsQL-MFS child scores were weak to moderate, although they were also statistically significant in most cases (the exception being the correlation with sleep/rest scores). In contrast, we anticipated somewhat weaker positive correlations between VFS and CDI-2 scores, given our underlying assumption that the VFS-Peds measures a different construct than depression. While this was observed in some cases, in others, it was not. For example, all correlations between VFS-$P_{Physical}$ and CDI-2 scores were weak in magnitude, and only the correlation with CDI-2, which is emotional scores, remained significant following Bonferroni adjustments. A similar weak, but statistically significant, association was observed between VFS-T and CDI-2: functional scores. However, correlations between VFS-P_{Mental} and all CDI-2 scores were statistically significant and moderate in magnitude. Likewise, associations between VFS-T and CDI-2 total and emotional scores and the VFS-C and CDI-2 total score were significant, and they were moderate in magnitude. These findings were inconsistent with our initial hypotheses and are discussed later.

Test Reliability

Results of analyses of test–retest reliability are shown in Table 7. The median length of time between test and retest measures was 5 days for the child and parent

Table 6. Spearman's rank correlations (r_s) between VFS scale scores (IRT scale scores) and PedsQL-MFS and CDI-2 scores.

	PedsQL-MFS				CDI-2		
Group	Total	General	Cognitive	Sleep/rest	Total	Emotional	Functional
*VFS-P_{Mental}	**−.73 (< .001)**	**−.59 (< .001)**	**−.72 (< .001)**	**−.38 (.006)**	**.63 (< .001)**	**.56 (< .001)**	**.60 (< .001)**
*VFS-$P_{Physical}$	**−.63 (< .001)**	**−.51 (< .001)**	**−.45 (.001)**	**−.55 (< .001)**	.35 (.013)	**.40 (.004)**	.25 (.079)
**VFS-C	**−.39 (.005)**	**−.41 (.003)**	**−.38 (.006)**	−.16 (.251)	**.42 (.002)**	NA	NA
***VFS-T	NA	NA	NA	NA	**.51 (< .001)**	**.54 (< .001)**	**.36 (.012)**

Note. *p* values (two-tailed) are shown in parentheses. Bolded values are significant after using a Bonferroni correction to adjust for multiple comparisons. The number of asterisks (*) corresponds to the corrected *p* values: $*p < .0071$; $**p < .01$; $***p < .0167$. VFS = Vanderbilt Fatigue Scale; IRT = item response theory; PedsQL-MFS = Pediatric Quality of Life Inventory Multidimensional Fatigue Scale; CDI-2 = Children's Depression Inventory; VFS-P = VFS parent proxy–report scale; VFS-C = VFS child self-report scale; NA = category is not applicable to respondents; VFS-T = VFS teacher proxy–report scale.

Table 7. Mean VFS summed scores and *SE*s at Time 1 and Time 2 for the Phase 4 respondents who completed the retest measure.

Summed Scores	*N*	Time 1	Time 2	Wilcoxon *Z*	Spearman's Rho	ICC
VFS-P$_{\text{Mental}}$	139	12.4 (0.65)	12.6 (0.60)	−1.06 (.291)	**.90 (< 0.001)**	**.90 [.87, .93]**
VFS-P$_{\text{Physical}}$	139	11.5 (0.42)	10.7 (0.41)	**−2.56 (.011)**	**.84 (< 0.001)**	**.84 [.78, .88]**
VFS-C	43	17.1 (1.47)	16.3 (1.60)	−.985 (.325)	**.79 (< 0.001)**	**.84 [.72, .91]**
VFS-T	73	19.0 (0.76)	18.9 (0.68)	−.458 (.647)	**.69 (< 0.001)**	**.72 [.58, .81]**

Note. Wilcoxon *Z* and the resultant *p* values in parentheses are shown in the fifth column. Spearman's correlation (r_s) and ICC values for comparisons of scores at T1 and T2 are shown in the sixth and final columns, respectively. ICC 95% confidence intervals are shown in brackets. Bolded values are significant. VFS = Vanderbilt Fatigue Scale; *SEs* = standard errors; *N* = number of participants in the sample; ICC = intraclass correlation coefficient; VFS-P = VFS parent proxy–report scale; VFS-C = VFS child self-report scale; VFS-T = VFS teacher proxy–report scale.

respondents (range: 5–32 days) and 13 days for teacher respondents (range: 6–36 days). A series of Wilcoxon signed-ranks tests were used to examine differences in mean total (child and teacher data) and subscale (parent data) scores obtained at baseline (T1) and approximately 1–2 weeks later (T2). Mean T1–T2 differences in summed scores were all quite small (i.e., < 1 point), and in general, the differences were not statistically significant. The one exception was the VFS-P$_{\text{Physical}}$ T1 and T2 data, which showed a small (0.8 point) but statistically significant decrease in fatigue at T2. The effect size for this difference, however, was small ($r = -.15$). In addition, we assessed temporal stability by examining correlations between individual T1 and T2 scores across scales. Spearman's correlation coefficients (r_s) were all positive, statistically significant, and strong in magnitude—ranging from .69 to .90. Together, these results suggest that the VFS-Peds have good, short-term, test–retest reliability (see Table 7).

Discussion

Many CHL are at increased risk for listening-related fatigue and its negative effects (e.g., Bess & Hornsby, 2014; Bess et al., 2020). Thus, there is a need for a sensitive, child-centered tool designed to identify and monitor those CHL who are experiencing this type of fatigue. In this article, we describe the development process for the VFS-Peds and report on their psychometric properties. Measuring a subjective experience like fatigue is inherently challenging, particularly in children; hence, multiple phases were employed to enhance the relevance and quality of our scales. This systematic development process, grounded in IRT, included generation and validation of relevant items based on input from target respondents (i.e., children, parents, and professionals); expert review; advanced statistical methods (IRT and DIF analyses) to assess item and test quality; and a systematic evaluation of the reliability and validity of the scales. This process yielded a set of short, psychometrically rigorous, reliable, and valid scales for assessing listening-related fatigue in children.

The VFS-Peds also appear to be user-friendly and clinically practical. Although additional validation is needed, it has been our experience that children 11 years of age and older are able to complete the self-report scale quickly (e.g., typically within 5 min) and easily with minimal adult guidance. However, children younger than this, or those with additional disabilities, may benefit from having an adult read the scale aloud and query them on their responses. Our informal experiences with adult proxies (parents and professionals) also revealed they were able to complete the scales quickly and independently. The relative lack of missing data (< 1% of item responses across all scales) supports the clarity of the scale items and their ease of use.

Related to clinical utility, the VFS-Peds can be scored using IRT methods or by summing item responses. To evaluate the sensitivity and precision of the VFS-Peds, we used IRT scale scores, derived using the EAP scoring method to analyze item response patterns, rather than using summed scores. We did this because IRT pattern scoring is more sensitive to differences among participants in the underlying latent construct (i.e., listening-related fatigue). However, IRT analysis of response patterns requires the use of a computer program and may not be feasible in all clinical settings. In these cases, calculating VFS-Peds summed scores provides a simple alternative for clinical use. For those wishing to compare summed scores to IRT scale scores, we also employed EAP on the summed score data (Lord & Wingersky, 1984) and provide conversion tables that relate summed scores to IRT scale scores (see Supplemental Material S6). Note that the EAP for a given summed score is calculated as an average IRT scale score over all possible response patterns. In addition, for those interested, we provide R code that can be used to calculate IRT scale scores based on VFS-Peds response data. The item discrimination and threshold estimates (see Supplemental Material S5) used in this code are provided as known parameters. The R code is freely available online (https://osf.io/vpjf5/).

From a theoretical perspective, one finding of interest from this study relates to the factor structure of listening-related fatigue. It is clear that feelings of fatigue can be elicited in a variety of different ways (e.g., a sprinter struggling to sustain their optimal pace vs. a student taking an extended exam). However, the extent to which these distinct causes of fatigue give rise to substantively different subjective experiences of fatigue remains unclear. In other words, the dimensionality of fatigue remains an empirical question and an active matter of debate (Lai et al., 2006; Michielsen et al., 2004; Smets et al., 1995). Adding to this debate, our findings suggest that the factor structure of listening-related fatigue may vary across respondent groups. Our initial literature review and qualitative findings from focus groups and interviews suggested that listening-related fatigue in CHL was multidimensional in nature, including physical, cognitive, and social–emotional characteristics (Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021). This finding prompted us to create test items for each scale that specifically targeted these various domains during the development phase. Despite this approach, results from our initial EFAs (with 60 items) suggested that the diverse expressions of listening-related fatigue, as self-reported by CHL, were well represented by a unidimensional construct. While results of initial EFAs based on responses from teachers and school professionals were not as clearcut as the child data, they also yielded a unique set of items consistent with a single underlying unidimensional construct. In contrast, EFA analyses of data from parent proxy respondents suggested that the listening-related fatigue experienced by children was multidimensional and reasonably described by a two-factor model including a "mental" factor (based on social, emotional, and cognitive items) and a physical factor with unique item sets for each factor.

The reason for the differing conclusions (unidimensional vs. multidimensional listening-related fatigue) based on respondent type is unclear and requires further study. One plausible explanation is that listening-related fatigue is truly multidimensional; however, parents are better able to observe, and thus report on, the physical sequelae of listening-related fatigue in their children (e.g., falling asleep right after school) relative to teachers and school professionals. This could explain why our parent proxy–report data suggested a distinct physical component to listening-related fatigue, whereas our teacher proxy–report data did not. The unidimensional nature of listening-related fatigue revealed in our child self-report data may reflect developmental limitations of children compared to adults. For example, Davis, Schlundt, Bonnet, Camarata, Hornsby, and Bess (2021) found that some CHL had difficulty defining their fatigue experiences and connecting it to their life experiences. When initially queried about their experiences with listening-related fatigue, some children suggested it was not a problem for them. However, with follow-up questioning, it was not uncommon for children to report exhibiting various fatigue-related behaviors (e.g., regular need for naps after school). Likewise, for some children who did not report fatigue-related issues, their parents would readily identify behaviors suggestive of such fatigue.

Another finding of interest was the stronger than expected association between listening-related fatigue (VFS-Peds scores) and depression (CDI-2 scores). Our initial hypothesis was that we would see the strongest associations between the VFS-Peds and the PedsQL-MFS, given both are measures of fatigue. We expected that associations with depression would be somewhat weaker, given that depression and fatigue are considered by some as distinct constructs. Prior work in adults supports this hypothesis. For example, De Vries et al. (2004) assessed fatigue and depression in a large sample of adults suffering from sarcoidosis. Fatigue was assessed using the FAS (Michielsen et al., 2003), and depression was measured using the Beck Depression Inventory (Beck et al., 1961). Results of a principal components analysis strongly suggested that fatigue and depression were separate factors. Michielsen et al. (2003) reported a similar finding. In contrast to these expectations, our results in children revealed a stronger than expected association between listening-related fatigue and depression and one that varied somewhat based on respondent type (parents vs. children and parent proxy–report subscales; see Table 6). For example, VFS-P_{Mental} scores yielded moderate-to-strong and significant associations with both PedsQL-MFS and CDI-2 total scores. Although the correlations were weaker, we saw a similar pattern for the child self-report data. Only the VFS-$\text{P}_{\text{Physical}}$ data showed the hypothesized weaker correlation with CDI-2 data.

Despite our initial predictions, many well-established unique constructs can show moderate correlations. For example, the cognitive construct of fluid reasoning, which is defined as a thinking ability, shows correlations of .30 to .45 with comprehension knowledge, which is defined as a measure of stored knowledge that is not measured using active problem-solving items (Schrank, 2011). Yet, it is well-established that fluid reasoning and comprehension knowledge are different broad abilities (McGrew, 2005). Likewise, in the fatigue literature, some research suggests that fatigue and depression are related but different constructs. This is consistent with the finding that some sequelae of fatigue (e.g., social isolation) are also seen in depression (Jacobsen et al., 2003). Supporting this hypothesis, Wilson et al. (2018) used confirmatory factor analyses to examine associations between fatigue and depression in postpartum women. Their results suggested that fatigue and depression were separate but related constructs that were moderately correlated (r = .41). It is also possible that the age of our participants may have played a role in our findings. For example, studies examining associations between fatigue and depression in children and adolescents who were undergoing treatment for cancer have observed significant associations between the constructs (Hinds et al., 2007; Hockenberry et al., 2003). Additional work in this area is required to better understand associations between listening-related fatigue and depression in children and adults.

Limitations and Future Directions

Our psychometric analyses indicate that the VFS-Peds are psychometrically reliable, valid, and clinically practical instruments that can be used to reliably measure listening-related fatigue over a wide range of severities. Despite the strong psychometric findings, the study has limitations worth noting. For example, cognitive interviews and our direct experiences with children suggest that all scale items are clear and understandable for our target respondents. However, we did not directly assess the reading or vision abilities of our respondents. Additional work examining the impact of literacy level and additional disabilities on the usability of the scales is needed to ensure that the scales are accessible to a wide audience. This additional work may be especially important when assessing young children using a self-report format. While children as young as 6 years of age provided responses via self-report, the vast majority (~92%) of our final (Phase 4) sample were ≥ 11 years old. Thus, additional work is needed to confirm the usability of the scale for this younger population. Additional tools, such as pictorial descriptors, may prove useful for very young children (e.g., 5–7 years old). The age distribution of the children reported on by our parents and teachers was much more uniform. Children as young as 6 were again included; however, ~53% and ~ 45% of children reported on by the parent and teacher groups, respectively, were ≤ 11 years old.

In addition, the clinical and practical applications of the scales require further development and study. For example, additional research is needed to better understand the "real-life" impact of listening-related fatigue on classroom behavior, learning, and quality of life. There are strong theoretical reasons to suggest that high levels of listening-related fatigue will have negative functional effects (e.g., see Camarata et al., 2018; Davis, Schlundt, Bonnet, Camarata, Hornsby, & Bess, 2021), but this must be tested empirically to determine appropriate cutpoint scores for targeted interventions. In addition, if elevated listening-related fatigue does impair children's listening and learning, future study will be required to develop and test effective interventions and/or compensatory strategies to address these impairments.

Conclusions

In summary, many CHL experience listening-related fatigue and its potential consequences prompting the need for a psychometrically sensitive and valid tool to identify the presence and severity of this type of fatigue. This article describes the development and validation of a suite of such scales: the VFS-Peds. The VFS-Peds provide a sensitive, reliable, and valid measure of listening-related fatigue in children that may be appropriate for clinical use. To provide content validity, we drew upon the complex experiences of a large sample of CHL and their parents, and teachers/school professionals. We employed advanced and robust statistical techniques (i.e., IRT and DIF analyses) to create scales with optimal precision. We hope that the VFS-Peds will be useful in future work to better understand the dimensionality and pervasiveness of listening-related fatigue in diverse pediatric populations (e.g., hearing loss and other disabilities and second language learners). Finally, we believe that these scales will provide important clinical value by capturing some of the broader implications of hearing loss not readily discernible using current diagnostic tests and serve as an additional marker of intervention success. Copies of the completed scales, which include instructions and scoring information, are available as supplemental files to this article (Supplemental Materials S2, S3, and S4). In addition, current and updated information related to the VFS-Peds (e.g., translated versions) and downloadable versions of the scales are also available at the following website: https://www.vumc.org/vfs/.

Author Contributions

Benjamin W. Y. Hornsby: Conceptualization (Supporting), Data curation (Supporting), Formal analysis (Supporting), Funding acquisition (Supporting), Investigation (Equal), Methodology (Supporting), Project administration (Supporting), Writing – original draft (Lead), Writing – review & editing (Lead). **Stephen Camarata:** Conceptualization (Supporting), Funding acquisition (Supporting), Investigation (Equal), Methodology (Equal), Project administration (Supporting), Writing – original draft (Supporting), Writing – review & editing (Equal). **Sun-Joo Cho:** Formal analysis (Lead), Methodology (Equal), Writing – review & editing (Supporting). **Hilary Davis:** Data curation (Lead), Formal analysis (Supporting), Investigation (Equal), Project administration (Supporting), Writing – review & editing (Equal). **Ronan McGarrigle:** Data curation (Supporting), Formal analysis (Supporting), Investigation (Supporting), Methodology (Supporting), Writing – review & editing (Supporting). **Fred H. Bess:** Conceptualization (Lead), Funding acquisition (Lead), Investigation (Equal), Methodology (Equal), Project administration (Lead), Writing – original draft (Supporting), Writing – review & editing (Equal).

Data Availability Statement

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Acknowledgments

The primary support for this work was provided by a grant from the Institute for Education Sciences (R324A160300; PI: Bess). Additional support was provided by a Vanderbilt Institute for Clinical and Translational Research Grant (UL1 TR000445 from NCATS/NIH) to the Vanderbilt University Medical Center. Graphic services were supported in part by EKS NICHD Grant 1P50HD103537-01 to the Vanderbilt Kennedy Center. The opinions expressed are those of the authors and do not represent the views of the U.S. Department of Education or the National Institutes of Health or other institutions. The authors would like to thank the families, professionals, and Vanderbilt University Medical Center graduate students who participated in each phase of the study. In addition, throughout the project, the authors received invaluable insight and feedback from multiple individuals with expertise in this area, including Cheryl DeConde Johnson, Mary Pat Moeller, Christine Yoshinaga-Itano, David Schlundt, Krystal Werfel, Emily Lund, Douglas Sladen, Andrea Dunn, Marc Hayes, and Kemberlee Bonnet. They would also like to thank Hope Lancaster for her thoughtful and helpful assistance with select statistical issues. They greatly appreciate the contributions of these individuals, without which this work would not have been possible.

References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bess, F. H., Davis, H., Camarata, S., & Hornsby, B. W. (2020). Listening-related fatigue in children with unilateral hearing loss. *Language, Speech, and Hearing Services in Schools, 51*(1), 84–97. https://doi.org/10.1044/2019_LSHSS-OCHL-19-0017

Bess, F. H., Dodd-Murphy, J., & Parker, R. A. (1998). Children with minimal sensorineural hearing loss: Prevalence, educational performance, and functional status. *Ear and Hearing, 19*(5), 339–354. https://doi.org/10.1097/00003446-199810000-00001

Bess, F. H., Gustafson, S. J., Corbett, B. A., Lambert, E. W., Camarata, S. M., & Hornsby, B. W. (2016). Salivary cortisol profiles of children with hearing loss. *Ear and Hearing, 37*(3), 334–344. https://doi.org/10.1097/AUD.0000000000000256

Bess, F. H., Gustafson, S. J., & **Hornsby, B. W.** (2014). How hard can it be to listen? Fatigue in school-age children with hearing loss. *Journal of Educational Audiology, 20,* 1–14.

Bess, F. H., & Hornsby, B. W. (2014). Commentary: Listening can be exhausting–fatigue in children and adults with hearing loss. *Ear and Hearing, 35*(6), 592–599. https://doi.org/10.1097/AUD.0000000000000099

Camarata, S., Werfel, K., Davis, T., Hornsby, B. W., & Bess, F. H. (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. https://doi.org/10.1177/0014402918773316

Choi, S. W., Gibbons, L. E., & Crane, P. K. (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), 1–30. https://doi.org/10.18637/jss.v039.i08

Christodoulou, C. (2007). The assessment and measurement of fatigue. In J. Deluca (Ed.), *Fatigue as a window to the brain* (pp. 19–35). MIT Press.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cole, D. A., Cho, S.-J., Martin, N. C., Youngstrom, E. A., March, J. S., Findling, R. L., Compas, B. E., Goodyer, I. M., Rohde, P., Weissman, M., Essex, M. J., Hyde, J. S., Curry, J. F., Forehand, R., Slattery, M. J., Felton, J. W., & Maxwell, M. A. (2012). Are increased weight and appetite useful indicators of depression in children and adolescents? *Journal of Abnormal Psychology, 121*(4), 838–851. https://doi.org/10.1037/a0028175

Davis, H., Schlundt, D., Bonnet, K., Camarata, S., Bess, F. H., & Hornsby, B. W. (2021). Understanding listening-related fatigue: Perspectives of adults with hearing loss. *International Journal of Audiology, 60*(6), 458–468. https://doi.org/10.1080/14992027.2020.1834631

Davis, H., Schlundt, D., Bonnet, K., Camarata, S., Hornsby, B. W., & Bess, F. H. (2021). Listening-related fatigue in children with hearing loss: Perspectives of children, parents, and school professionals. *American Journal of Audiology, 30*(4), 929–940. https://doi.org/10.1044/2021_AJA-20-00216

De Ayala, R. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*(2), 155–170. https://doi.org/10.1177/014662169401800205

Deluca, J. (Ed.). (2005). *Fatigue as a window to the brain.* MIT Press. https://doi.org/10.7551/mitpress/2967.001.0001

De Vries, J., Michielsen, H., Van Heck, G. L., & Drent, M. (2004). Measuring fatigue in sarcoidosis: the Fatigue Assessment Scale (FAS). *British Journal of Health Psychology, 9*(3), 279–291. https://doi.org/10.1348/1359107041557048

Dwyer, R. T., Gifford, R. H., Bess, F. H., Dorman, M., Spahr, A., & Hornsby, B. W. Y. (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. https://doi.org/10.1044/2019_AJA-19-0009

Eiser, C., & Morse, R. (2001). Can parents rate their child's health-related quality of life? Results of a systematic review. *Quality of Life Research, 10*(4), 347–357. https://doi.org/10.1023/A:1012253723272

Gaba, D. M., & Howard, S. K. (2002). Fatigue among clinicians and the safety of patients. *New England Journal of Medicine, 347*(16), 1249–1255. https://doi.org/10.1056/NEJMsa020846

Gustafson, S. J., Key, A. P., Hornsby, B. W. Y., & Bess, F. H. (2018). Fatigue related to speech processing in children with hearing loss: Behavioral, subjective, and electrophysiological measures. *Journal of Speech, Language, and Hearing Research, 61*(4), 1000–1011. https://doi.org/10.1044/2018_JSLHR-H-17-0314

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science + Business Media, LLC.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research, 45*(3), 573–584. https://doi.org/10.1044/1092-4388(2002/046)

Hinds, P. S., Hockenberry, M., Tong, X., Rai, S. N., Gattuso, J. S., McCarthy, K., Pui, C. H., & Srivastava, D. K. (2007). Validity and reliability of a new instrument to measure cancer-related fatigue in adolescents. *Journal of Pain and Symptom Management, 34*(6), 607–618. https://doi.org/10.1016/j.jpainsymman.2007.01.009

Hockenberry, M. J., Hinds, P. S., Barrera, P., Bryant, R., Adams-McNeill, J., Hooke, C., Rasco-Baggott, C., Patterson-Kelly, K., Gattuso, J. S., & Manteuffel, B. (2003). Three instruments to assess fatigue in children with cancer: The child, parent and staff perspectives. *Journal of Pain and Symptom Management, 25*(4), 319–328. https://doi.org/10.1016/S0885-3924(02)00680-2

Hockenberry-Eaton, M., Hinds, P., Howard, V., Gattuso, J., O'Neill, J., Alcoser, P., Bottomley, S., Kline, N., & Euell, K. (1999). Developing a conceptual model for fatigue in children. *European Journal of Oncology Nursing, 3*(1), 5–11. https://doi.org/10.1016/1462-3889(91)80005-7

Hockey, R. (2013). *The psychology of fatigue: Work, effort and control* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139015394

Holman, J. A., Drummond, A., Hughes, S. E., & Naylor, G. (2019). Hearing impairment and daily-life fatigue: A qualitative study. *International Journal of Audiology, 58*(7), 408–416. https://doi.org/10.1080/14992027.2019.1597284

Holman, J. A., Drummond, A., & Naylor, G. (2021). Hearing aids reduce daily-life fatigue and increase social activity: A longitudinal study. *Trends in Hearing, 25*. https://doi.org/10.1177/23312165211052786

Holman, J. A., Hornsby, B. W. Y., Bess, F. H., & Naylor, G. (2021). Can listening-related fatigue influence well-being? Examining associations between hearing loss, fatigue, activity levels and well-being. *International Journal of Audiology, 60*(Suppl. 2), 47–59, https://doi.org/10.1080/14992027.2020.1853261

Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing, 34*(5), 523–534. https://doi.org/10.1097/AUD.0b013e31828003d8

Hornsby, B. W. Y., Camarata, S., Cho, S.-J., Davis, H., McGarrigle, R., & Bess, F. H. (2021). Development and validation of the Vanderbilt Fatigue Scale for Adults (VFS-A). *Psychological Assessment, 33*(8), 777–788. https://doi.org/10.1037/pas0001021

Hornsby, B. W. Y., Gustafson, S. J., Lancaster, H., Cho, S.-J., Camarata, S., & Bess, F. H. (2017). Subjective fatigue in children with hearing loss assessed using self- and parent-proxy report. *American Journal of Audiology, 26*(3S), 393–407. https://doi.org/10.1044/2017_AJA-17-0007

Hornsby, B. W. Y., & Kipp, A. M. (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. https://doi.org/10.1097/AUD.0000000000000203

Hornsby, B. W. Y., Naylor, G., & Bess, F. H. (2016). A taxonomy of fatigue concepts and their relation to hearing loss. *Ear and Hearing, 37*(Suppl. 1), 136S–44S. https://doi.org/10.1097/AUD.0000000000000289

Hornsby, B. W. Y., Werfel, K., Camarata, S., & Bess, F. H. (2014). Subjective fatigue in children with hearing loss: Some preliminary findings. *American Journal of Audiology, 23*(1), 129–134. https://doi.org/10.1044/1059-0889(2013/13-0017)

Hospers, J. M. B., Smits, N., Smits, C., Stam, M., Terwee, C. B., & Kramer, S. E. (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. https://doi.org/10.1044/2015_JSLHR-H-15-0156

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

Jacobsen, P. B., Donovan, K. A., & Weitzner, M. A. (2003). Distinguishing fatigue and depression in patients with cancer. *Seminars in Clinical Neuropsychiatry, 8*(4), 229–240.

Johnson, S. K. (2005). Depression and fatigue. In J. DeLuca (Ed.), *Fatigue as a window to the brain* (pp. 157–172). MIT Press.

Kang, T., & Chen, T. T. (2008). Performance of the GeneralizedS-X2Item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kovacs, M., & Multi-Health Systems Staff. (2011). *Children's Depression Inventory 2nd Edition (CDI 2): Technical manual*. Multi-Health Systems.

Lai, J. S., Crane, P. K., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Quality of Life Research, 15*(7), 1179–1190. https://doi.org/10.1007/s11136-006-0060-6

Lewis, D., Schmid, K., O'Leary, S., Spalding, J., Heinrichs-Graham, E., & High, R. (2016). Effects of noise on speech recognition and listening effort in children with normal hearing and children with mild bilateral or unilateral hearing loss. *Journal of Speech, Language, and Hearing Research, 59*(5), 1218–1232. https://doi.org/10.1044/2016_JSLHR-H-15-0207

Lindburg, M., Ead, B., Jeffe, D. B., & Lieu, J. E. C. (2021). Hearing loss-related issues affecting quality of life in preschool children. *Otolaryngology–Head and Neck Surgery, 164*(6), 1322–1329. https://doi.org/10.1177/0194599820962475

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*(4), 453–461. https://doi.org/10.1177/014662168400800409

McGarrigle, R., Gustafson, S. J., Hornsby, B. W. Y., & Bess, F. H. (2019). Behavioral measures of listening effort in school-age children: Examining the effects of signal-to-noise ratio, hearing loss, and amplification. *Ear and Hearing, 40*(2), 381–392. https://doi.org/10.1097/AUD.0000000000000623

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan &

P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136–181). The Guilford Press.

Michielsen, H. J., De Vries, J., & Van Heck, G. L. (2003). Psychometric qualities of a brief self-rated fatigue measure: The fatigue assessment scale. *Journal of Psychometric Research, 54*(4), 345–352. https://doi.org/10.1016/S0022-3999(02)00392-6

Michielsen, H. J., De Vries, J., Van Heck, G. L., Van de Vijver, F. J., & Sijtsma, K. (2004). Examination of the dimensionality of fatigue. *European Journal of Psychological Assessment, 20*(1), 39–48. https://doi.org/10.1027/1015-5759.20.1.39

Muthen, L., & Muthen, B. (1998–2017). *Mplus user's guide* (8th ed.).

Patrick, D. L., & Deyo, R. A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care, 27*(Suppl. 3), S217–S232. https://doi.org/10.1097/00005650-198903001-00018

Ravid, S., Afek, I., Suraiya, S., Shahar, E., & Pillar, G. (2009). Kindergarten children's failure to qualify for first grade could result from sleep disturbances. *Journal of Child Neurology, 24*(7), 816–822. https://doi.org/10.1177/0883073808330766

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(S1), 1–97. http://doi.org/10.1007/BF03372160

Schrank, F. A. (2011). Woodcock-Johnson III Tests of Cognitive Abilities. In A. S. Davis (Ed.), *Handbook of pediatric neuropsychology* (pp. 415–434). Springer.

Sindhar, S., Friesen, T. L., Carpenter, D., Kesser, B., & Lieu, J. (2021). Fatigue in children with unilateral and bilateral hearing loss. *Otology & Neurotology, 42*(9), e1301–e1307. https://doi.org/10.1097/MAO.0000000000003225

Smets, E. M., Garssen, B., Bonke, B. D., & De Haes, J. C. (1995). The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research, 39*(3), 315–325. https://doi.org/10.1016/0022-3999(94)00125-o

Sprangers, M. A., & Aaronson, N. K. (1992). The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. *Journal of Clinical Epidemiology, 45*(7), 743–760. https://doi.org/10.1016/0895-4356(92)90052-o

Steiger, J. H., & 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, K. D., Jacobsen, P. B., Blanchard, C. M., & Thors, C. (2004). Further validation of the multidimensional fatigue symptom inventory-short form. *Journal of Pain and Symptom Management, 27*(1), 14–23. https://doi.org/10.1016/j.jpainsymman.2003.06.003

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. https://dxdoi.org/10.1007/BF02291170

Varni, J. W., Burwinkle, T. M., Katz, E. R., Meeske, K., & Dickinson, P. (2002). The PedsQL™ in pediatric cancer. *Cancer, 94*(7), 2090–2106. https://doi.org/10.1002/cncr.10428

Varni, J. W., Limbers, C. A., Bryant, W. P., & Wilson, D. P. (2010). The PedsQL™ Multidimensional Fatigue Scale in pediatric obesity: Feasibility, reliability and validity. *International Journal of Pediatric Obesity, 5*(1), 34–42. https://doi.org/10.3109/17477160903111706

Varni, J. W., Limbers, C. A., & Burwinkle, T. M. (2007). Literature review: Health-related quality of life measurement in pediatric oncology: Hearing the voices of the children. *Journal of Pediatric Psychology, 32*(9), 1151–1163. https://doi.org/10.1093/jpepsy/jsm008

Werfel, K. L., & Hendricks, A. E. (2016). The relation between child versus parent report of chronic fatigue and language/literacy skills in school-age children with cochlear implants. *Ear and Hearing, 37*(2), 216–224. https://doi.org/10.1097/AUD.0000000000000242

Wessely, S., Hotopf, M., & Sharpe, M. (1998). *Chronic fatigue and its syndromes* (1st ed.). Oxford University Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.

Wilson, N., Wynter, K., Fisher, J., & Bei, B. (2018). Related but different: Distinguishing postpartum depression and fatigue among women seeking help for unsettled infant behaviours. *BMC Psychiatry, 18*(1), 309. https://doi.org/10.1186/s12888-018-1892-7

Yu, C. Y. (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, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.