

UCSF

UC San Francisco Previously Published Works

Title

Using Item-Response Theory to Improve Interpretation of the Trans Woman Voice Questionnaire

Permalink

<https://escholarship.org/uc/item/5sk9w1qh>

Journal

The Laryngoscope, 133(5)

ISSN

0023-852X

Authors

Zhao, Nina W
Mason, James M
Blum, Alexander M
et al.

Publication Date





2023-05-01

DOI

10.1002/lary.30360

Peer reviewed

Using Item-Response Theory to Improve Interpretation of the Trans Woman Voice Questionnaire

Nina W. Zhao, MD, MAEd ; James M. Mason, MEd ; Alexander M. Blum, PhD; Eric K. Kim, BA; VyVy N. Young, MD ; Clark A. Rosen, MD ; Sarah L. Schneider, CCC-SLP

Objective: The Trans Woman Voice Questionnaire (TWVQ) is commonly used to quantify self-perceptions of voice for trans women seeking gender-affirming voice care, but the interpretation of TWVQ scores remains challenging. The objective of this study was to use item-response theory (IRT) to evaluate the relationship between TWVQ items and persons on a common scale and identify improvements to increase the meaningfulness of TWVQ scores.

Methods: A retrospective review of TWVQ scores from trans women patients between 2018–2020 was performed. Rasch-family models were used to generate item-person maps positioning respondent location and item difficulty estimates on a logit scale, which was then converted into a scaled score using linear transformations.

Results: TWVQ responses from 86 patients were analyzed. Initial item-person maps demonstrated that the middle response categories (“sometimes” and “often”) performed inconsistently across items (poor threshold banding); interpretability improved when these ratings were scored as one category. The models were rerun using revised scoring, which retained high reliability (0.93) and supported a unidimensional construct. Updated item-person maps revealed four scaled score zones (≤ 54 , >54 to ≤ 101 , >101 to ≤ 140 , and >140) that each corresponded to an increasing pattern of item thresholds (probability of selecting one response category vs. others). These ranges can be interpreted as minimal, low, moderate, and high, respectively.

Conclusions: Empiric data from Rasch analysis supports new interval scoring for the TWVQ that advances the clinical and research utility of the instrument and lays the foundation for future improvements in clinical care and outcomes assessment.

Key Words: instrument development, item-response theory, latent variable modeling, Rasch analysis, transgender voice.

Level of Evidence: NA

Laryngoscope, 00:1–8, 2022

From the Department of Otolaryngology – Head and Neck Surgery (N.W.Z., E.K.K., V.N.Y., C.A.R., S.L.S.), University of California, San Francisco, San Francisco, California, U.S.A.; Department of Otolaryngology – Head and Neck Surgery (N.W.Z.), University of California, Davis, Sacramento, California, U.S.A.; Graduate School of Education (J.M.M.), University of California, Berkeley, Berkeley, California, U.S.A.; Department of Special Education (A.M.B.), San Francisco State University, San Francisco, California, U.S.A.; Graduate School of Education (A.M.B.), Stanford University, Palo Alto, California, U.S.A.; Enrich Your Academics (A.M.B.), Emeryville, California, U.S.A.

Additional supporting information may be found in the online version of this article.

Present address: Nina W. Zhao, Department of Otolaryngology – Head and Neck Surgery, University Hospitals Cleveland Medical Center, Cleveland, Ohio, U.S.A.

Editor's Note: This Manuscript was accepted for publication on August 04, 2022.

This work was presented at the American Laryngological Association Annual Meeting, Dallas, Texas, U.S.A., April 2022.

A portion of this work was funded by the UCSF Academy of Medical Educators.

C.A.R. reports the following disclosures and financial relationships: Olympus America Inc, consultant; Instrumentarium, royalties; Freudenberg Medical, consultant; Reflux Gourmet LCC, shareholder. S.L.S. reports the following disclosures and financial relationships: MedBridge royalties. The other authors have no financial relationships or conflicts of interest to disclose.

The authors have no other funding, financial relationships, or conflicts of interest to disclose.

Send correspondence to Nina W. Zhao, MD, MAEd, Department of Otolaryngology – Head and Neck Surgery, University Hospitals Cleveland Medical Center, 11100 Euclid Ave, Cleveland, OH, 44106. E-mail: nina.zhao.md@gmail.com

DOI: 10.1002/lary.30360

INTRODUCTION

The Trans Woman Voice Questionnaire (TWVQ) is a 30-item patient self-report tool designed to measure trans women's perceptions of their voice and its impact on their everyday lives.¹ The TWVQ is often used by laryngologists and speech-language pathologists to supplement perceptual, acoustic, and other instrumental voice assessments during evaluations of transgender women seeking gender-affirming voice care. The TWVQ addresses the need for a more population-specific voice questionnaire and is developed from expert review with input from transgender individuals.^{1,2} Individuals complete TWVQ items using a four-point Likert scale (1 = never or rarely, 2 = sometimes, 3 = often, 4 = usually or always), and each item is summed to produce a total score with a minimum of 30 and maximum of 120. Previous studies demonstrated the TWVQ exhibits high reliability and established various sources of validity, including that scores inversely correlate with self-perception of voice femininity and improve after gender-affirming interventions.^{1,3–6}

Despite growing evidence for the validity of the TWVQ to assess the impact of voice on trans women's lives, several issues still hinder its usefulness in practice. First, although a lower score represents a more desirable outcome, the numbers alone lack meaning.^{4,7} For example, although 75 is in the middle of the scale, there is no clear data to support interpretation as a “moderate” impact. As a result, the nature of a patient's presentation

based on the raw score is unknown. Second, due to the ordinal nature of Likert-type responses, scores are not truly interval; a difference in raw score at one portion of the scale (e.g., 35–30) may not mean the same as an equivalent difference at another (e.g., 120–115). This issue leads to challenges in comparing treatment interventions for patients with differing TWVQ scores. Third, the dimensionality of the instrument has been inconsistent in the literature. Previously, Dacakis et al. reported a two-factor structure labeled as “vocal functioning” and “social participation.”⁴ However, a more recent study by Bultynck et al. described a three-factor structure: “anxiety and avoidance,” “vocal identity,” and “vocal function.”⁸ Although these factor organizations are similar, the items do not completely align between the two models, suggesting that the factors may not be distinct dimensions. Finally, there is currently limited understanding of how to compare results from different voice instruments. Prior work from this group showed that TWVQ and Voice Handicap Index-10 (VHI-10) scores moderately correlate,^{7,9} but additional insight into the relationship between scores would allow clinicians to make more informed instrument choices.

Much of these current issues stem from the methodologic limitations of classical test theory (CTT), which was the basis of prior TWVQ psychometric evaluations. Although CTT is widely used, there are multiple theoretical shortcomings, including the assumption of a linear relationship between observed scores and the variable of interest, the sample-dependent nature of the parameter estimates,^{10,11} and basing score meaning solely on the total score, requiring norm-referenced standards.¹² However, more contemporary approaches to measurement based on latent variable modeling known as item-response theory (IRT) may overcome these limitations. IRT is a family of statistical models that describes the relationship between latent constructs (e.g., a person or respondent’s “ability” or “trait”) and observed outcomes (e.g., responses to items). The latent trait is assumed to be organized in a continuum along a scale; therefore, the main goals of IRT are to estimate both person and item locations on that scale and describe the relationship between persons and items.^{10–17}

Unlike CTT, IRT focuses on item-level rather than test-level data to test assumptions and fit; parameter estimates are considered sample-independent as scores are referenced directly to the items.^{10–12,16} The models address the problem of ordinal rating scales by applying nonlinear transformations to create equal-interval log-odds units called logits. Both persons and items are subsequently calibrated on the same logit scale representing the construct of interest. Person and item locations along the construct are estimated from responses to individual items. Together, this information can be used to assess sources of measurement validity.^{16,18,19}

An item-level perspective using IRT can move beyond the limits of CTT and provide insight into how the TWVQ can quantify different levels of voice-related life impact. The goal of this study is to demonstrate the power and utility of IRT by using Rasch-family models to explore the relationship between TWVQ items and individuals’ responses, the dimensionality of the instrument,

and the relationship to the VHI-10. An updated interval scoring system supported by initial reliability and validity evidence is then proposed to increase the usefulness of TWVQ scores for clinical care and research.

MATERIALS AND METHODS

Study Design and Data Collection

A retrospective review of TWVQ scores from adult (≥18 years) trans women patients seeking treatment at a tertiary academic institution’s voice center between 2018–2020 was performed. All referrals for evaluation of “dysphonia” and “gender dysphoria” were reviewed. Responses to TWVQ items were individually recorded. Patient demographic information and relevant medical and surgical histories were also collected. The University of California, San Francisco Institutional Review Board approved this study (#18-26766).

Measurement Models

Three statistical measurement models were applied to the data, summarized below. All models were conducted on ConQuest software.²⁰

Model 1: Partial Credit Model

To place respondents and items on a common logit scale and examine internal structure, reliability, and item fit, the Rasch Partial Credit Model (PCM) was used. This ordinal IRT model estimates the locations of persons, items, and levels within item based on the maximum likelihood of observed response patterns.^{21,22} Persons higher along the continuum of gender-related voice impact have a higher probability of endorsing any given TWVQ item than persons at lower locations, and any given person has a higher probability of endorsing both less severe items and lower levels within items than more severe items or levels. Due to IRT conventions, item levels typically start at 0, so the software models responses “never or rarely,” “sometimes,” “often,” and “usually or always” as 0, 1, 2, and 3, respectively.²⁰

Item-person map. The results of the PCM were used to form an item-person map, also called a Wright Map, which is a visual representation of instrument structure depicting person locations and item thresholds along the construct, in logits.^{12,15,17,23,24} Persons higher along the scale possess more of the construct. Item response thresholds represent levels within items (e.g., never sometimes, often, always), and higher thresholds measure higher levels of the construct.^{15,17,23,24} As explained further in Results, these thresholds correspond to the location on the scale where respondents are equally likely to endorse the item at or above the level in question versus cumulative level(s) below. Placing persons and items on the same visual map allows for assessment of not only how well the two align, but also other patterns in responses, such as threshold banding. Although lower thresholds will always be below higher thresholds within an item, threshold banding occurs when the same thresholds across items are “banded” into distinct segments, which helps to divide the scale into descriptive regions based on response patterns.^{17,23,24}

Interval scaled score. To aid score interpretation, a linear transformation (scaled score = 20·logit + 145) was applied to the logit scale. The transformed scale preserves interval properties while avoiding negative values and decimals.

Item fit. Item fit indicates the degree to which observed item responses matches model predictions, reported as weighted mean square (WMNSQ). A reasonable WMNSQ is between 0.75 and 1.33.^{15,17} A WMNSQ less than 0.75 suggests the item

responses are more predictable than the model, whereas greater than 1.33 suggests responses are less predictable; thus, elevated values are more problematic than lower.

Mean location. Item-level internal structure can be examined through the mean person location at each response level within an item (average position of a person along the scale that endorses an item at a specific level).¹⁶ If an item is functioning properly, respondent mean locations would increase as item levels increase because, on average, persons responding at higher levels within an item should also respond higher to other items, locating themselves higher on the logit scale.

Reliability. Reliability was examined using internal consistency coefficients: Cronbach's alpha and person separation reliability, a Rasch model equivalent.^{16,17}

Model 2: Multidimensional PCM

To assess the dimensionality of the TWVQ, the Rasch multidimensional PCM^{25,26} was used to generate latent correlations between previously proposed factor structures of the TWVQ.^{4,8} The model assumes that each dimension is a separate latent variable and directly estimates the correlations between them. A high latent correlation between factors argues that the dimensions are better interpreted as a unidimensional construct, and vice versa.

Model 3: Latent Regression PCM

To examine relationship between the TWVQ and the VHI-10, the Rasch latent regression PCM was employed, which, by estimating the regression coefficients simultaneously with the measurement model, incorporates the error associated with estimating latent variables.²⁷ Using this model, TWVQ and VHI-10 scores were regressed and the output overlaid onto the same Wright Map, which allows assessment of how the two instruments compare in their measurement of voice-related life impact.

RESULTS

Patient Sample

Eighty-six patients were included in our sample (Table I). All patients were assigned male at birth. Eighty-five (98.9%) identified as female; one patient identified as non-binary, used she/her pronouns, and desired a more feminine voice so was included in the sample. Only 2 (2.7%) patients reported prior voice feminization surgery.

Initial Wright Map (Model 1)

An initial Wright Map (Fig. 1) was developed by fitting Model 1 to the original TWVQ. As described in Methods, the Wright Map is a visual representation of instrument structure, positioning mean respondent locations (vertical histogram), and item response thresholds (shaded bars) on a common logit scale (vertical axis). A person's location relative to an item's thresholds determines the likelihood of responding to that item a particular way. Persons higher on the map are likely to experience more voice-related life impact and endorse items at higher levels. Specifically, persons located at a particular item threshold have a 50% chance of endorsing the item at or above that threshold versus cumulative levels below. At the first threshold, the cumulative likelihood of answering "1" (sometimes) or above

TABLE I.
Patient Demographics.

Characteristic	
Assigned male at birth, <i>N</i> (%)	86 (100%)
Age, mean ± SD (years)	33.1 ± 11.2
Gender, <i>N</i> (%)	
Female	85 (98.9)
Non-binary	1 (1.2)
Duration of social transition, mean ± SD (mo) <i>N</i> = 65	41.8 ± 60.4
Duration of medical transition, mean ± SD (mo) <i>N</i> = 70	25.9 ± 34.8
Gender-affirming head and neck surgery, <i>N</i> (%)	
Facial feminization surgery	11 (15.1)
Voice feminization surgery	2 (2.7)
Tracheal shave/chondrolaryngoplasty	2 (2.7)
None	61 (83.6)
Unknown	13 (15.1)
Prior voice diagnoses	
Vocal fold atrophy	1 (1.2)
Muscle tension dysphonia	1 (1.2)
Smoking history	
Current	15 (17.4)
Former	9 (10.5)
Never	62 (72.1)
Stroboscopy, <i>N</i> (%)	
Normal	79 (91.9)
Abnormal	4 (4.7)
Not performed	3 (3.5)
Average TWVQ, mean ± SD (range, median, IQR)	87 ± 22 (35–120, 89, 34)
Average VHI-10, mean ± SD (range, median, IQR)	18 ± 10 (0–40, 16, 16.5)

IQR = interquartile range; SD = standard deviation; TWVQ = Trans Woman Voice Questionnaire; VHI-10 = Voice Handicap Index-10.

equals the likelihood of answering "0" (never or rarely). At the second threshold, the likelihood of answering "2" (often) or above equals the likelihood of answering any of the lower levels, and so on. Persons located above an item threshold have more than a 50% chance of endorsing the item at that threshold and those below have less than a 50% chance.

The Wright Map revealed that most of the respondents aligned with the item thresholds, with only a few individuals above the highest thresholds. However, threshold banding was poor. Specifically, Threshold 2 overlapped with the other two thresholds, suggesting that it may be less meaningful to score the middle categories "sometimes" and "often" differently. Therefore, the analysis was repeated with the original responses "0" (never or rarely), "1" (sometimes), "2" (often), and "3" (usually or always) modeled as 0, 1, 1, and 2, respectively. This new scoring was used for all subsequent analyses.

Revised Wright Map and Scaled Score (Model 1)

Updated scoring yields minimum and maximum sum scores of 0 and 60. After refitting Model 1 these

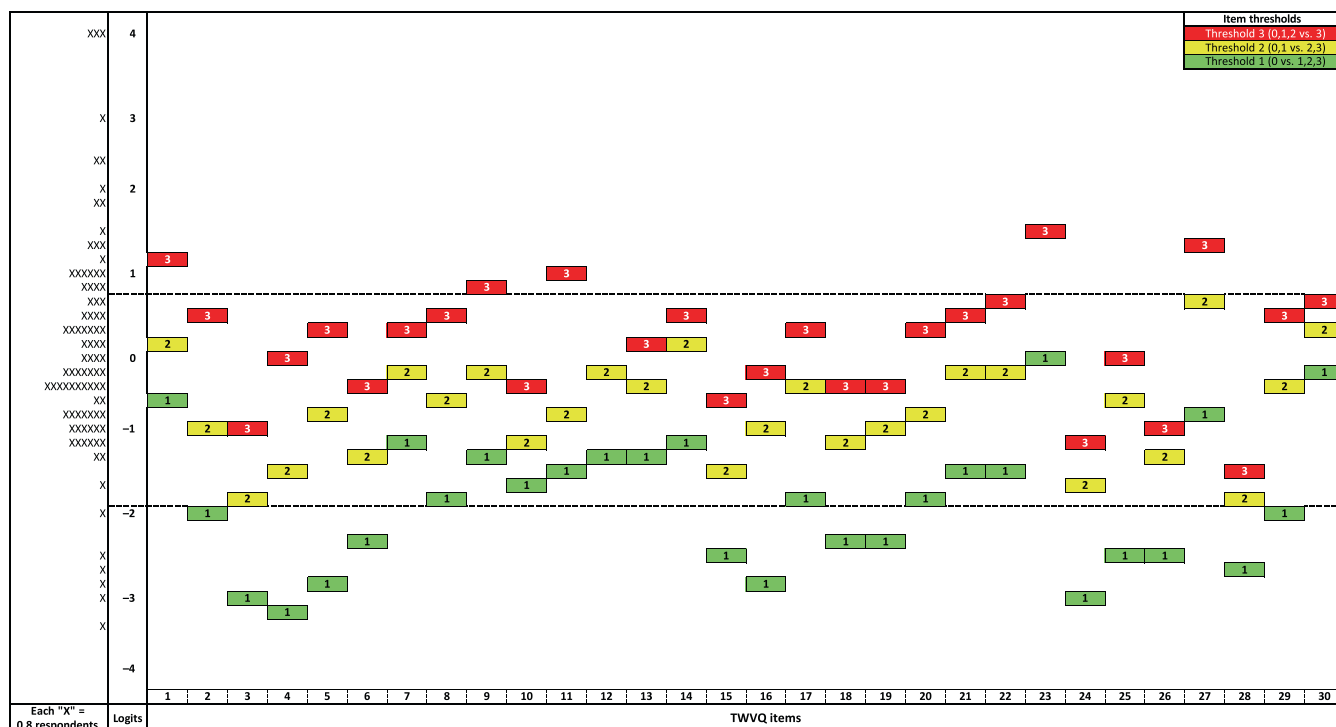


Fig. 1. Initial Wright Map depicting mean respondent locations and item response thresholds on a logit scale for the original Trans Woman Voice Questionnaire (TWVQ). Each “X” represents a group of 0.8 respondents. Numbered bars represent item response thresholds. The first threshold (1) shows the location on the map where respondents are 50% likely to respond “0” (never or rarely) and 50% likely to respond “1” (sometimes) or above. The second threshold (2) shows the location where respondents are 50% likely to respond “1” (sometimes) or below and 50% likely to respond “2” (often) or above. The third threshold (3) shows the location where respondents are 50% likely to respond “2” (often) or below and 50% likely to respond “3” (usually or always). Thick dashed lines depict the region of threshold 2, which entirely overlaps with regions of thresholds 1 and 3. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

scores corresponded to -7.21 and $+4.75$ logits, respectively, but note that the relationship between TWVQ score and the logit scale is not linear, particularly at the extremes (Fig. 2). The logit scale was then transformed into scaled scores ranging from 1 to 240. For complete score conversions, see Table S1.

As seen on the revised Wright Map (Fig. 3A), TWVQ sum scores of our sample ranged from 4 to 60 (person location -4.78 to $+4.75$ logits, scaled score 49–240) and item thresholds ranged from -4.51 to 2.17 logits. The map also revealed improved threshold banding as Threshold 1 for most items remained below all Threshold 2’s, demonstrating improved separation between

respondents who tend to respond “1” (sometimes or often) versus those who respond “2” (usually or always). Regions of differing response patterns also became visible, particularly when items were reordered by first thresholds (Fig. 3B); therefore, the Wright Map was divided into different interpretive zones. Respondents positioned at a scaled score 54 or below (Zone 0) were likely responding “0” (never or rarely) to nearly all items. Respondents above 54 and at or below 101 (Zone 1) were likely increasingly responding “1” (sometimes or often) to items with lower first thresholds. Respondents above 101 and at or below 140 were in a transition region (Zone 2) where they were likely responding “2” (usually or always) to more items with lower first thresholds while also increasingly responding “1” to items with higher first thresholds. Finally, respondents above 140 (Zone 3) are most likely increasingly responding “2” to all the items.

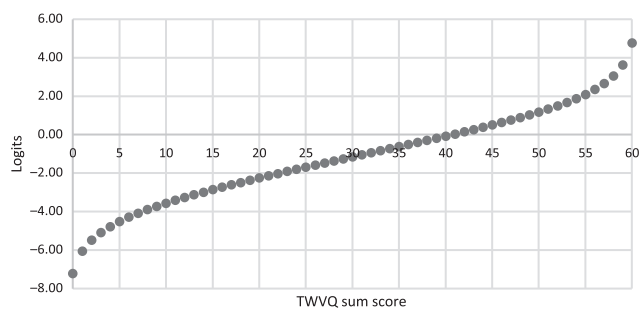


Fig. 2. Relationship between the updated sum score for the Trans Woman Voice Questionnaire (TWVQ) and the logit scale.

Item Fit and Mean Location (Model 1)

Nearly all items demonstrated good fit. Only three items, 1, 24, 28, were outside the desired WMNSQ range of 0.75–1.33 with values of 1.74, 0.70, and 1.40, respectively. Complete fit statistics are provided Table S2. Expected mean location also increased across all the TWVQ items (Table S3).

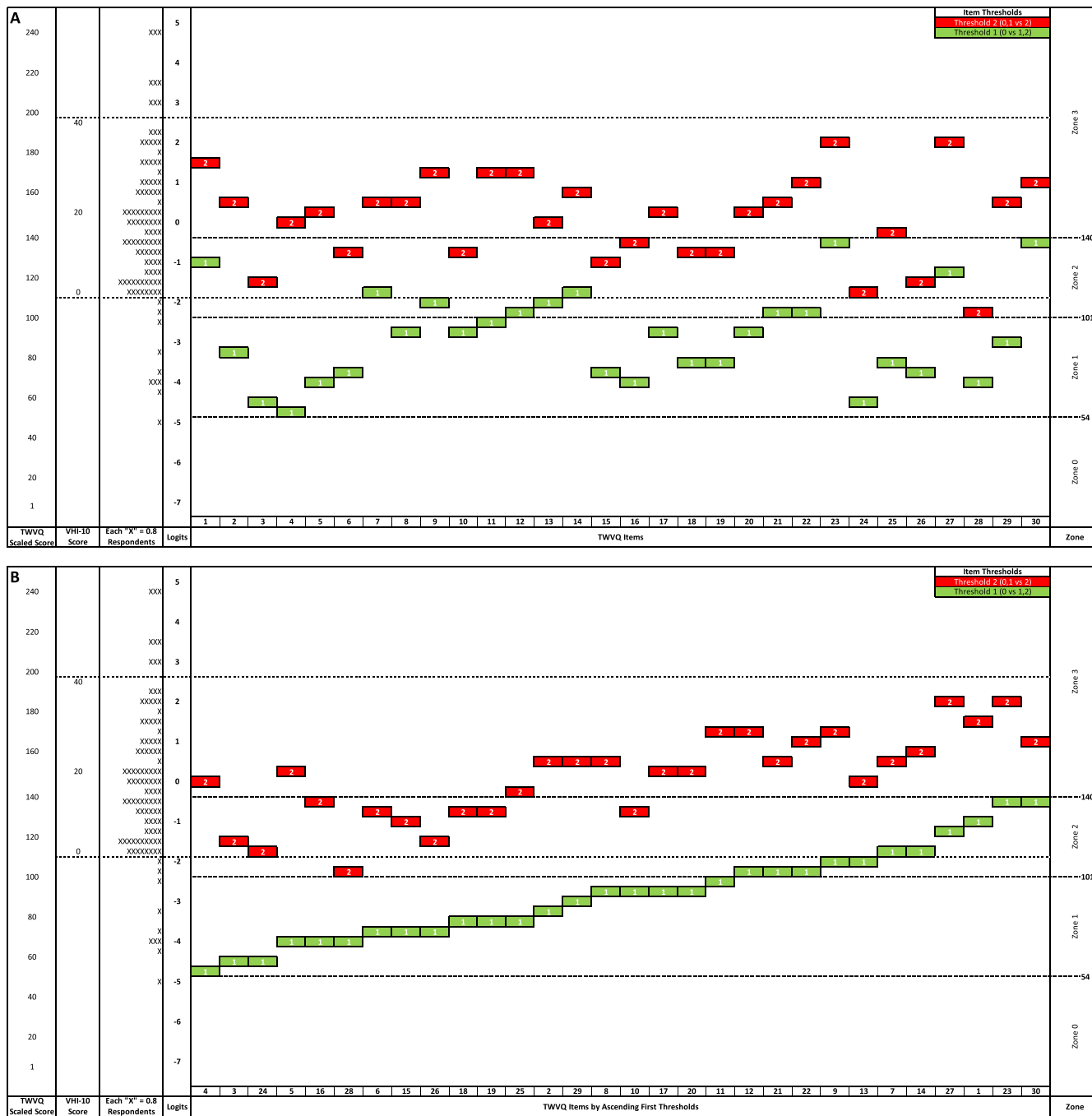


Fig. 3. Revised Wright Maps based on updated scoring (0 [never or rarely], 1 [sometimes], 1 [often], and 2 [usually or always]) and (A) ordered by item number or (B) ordered by ascending first level thresholds. Each “X” represents a group of 0.8 respondents. Numbered bars represent item response thresholds. The first threshold (1) shows the location on the map where respondents are 50% likely to respond “0” (never or rarely) and 50% likely to respond “1” (sometimes or often) or above. The second threshold (2) shows the location where respondents are 50% likely to respond “1” (sometimes or often) or below and 50% likely to respond “2” (usually or always). Thick dashed lines depict interpretive zones based on item response patterns: ≤54 (Zone 0), >54 to ≤101 (Zone 1), >101 to ≤140 (Zone 2), and >140 (Zone 3). Region between thin dotted lines shows the extent to which the VHI-10 overlaps with the Trans Woman Voice Questionnaire (TWVQ). VHI-10 = Voice Handicap Index-10. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

Reliability (Model 1)

Reliability of the updated scoring was high; Cronbach’s alpha and person separation reliability values were 0.95 and 0.93, respectively.

Dimensionality (Model 2)

Latent correlation coefficients corresponding to previously proposed TWVQ factor structures^{4,8} were positive and strong (Table II), indicating these factors are likely

TABLE II.
Latent Correlation Coefficient Tables for Previously Reported Two-Factor (A) and Three-Factor (B) Models.

(A) Two-Factor Model (Dacakis et al. ⁴)*			
Factor	1	2	
1. Vocal functioning	—		
2. Social participation	0.83	—	
(B) Three-Factor Model (Bultynck et al. ⁸)†			
Factor	1	2	3
1. Anxiety-avoidance	—		
2. Vocal identity	0.79	—	
3. Vocal functioning	0.87	0.81	—

*Dacakis et al.'s two-factor model consisted of a 14-item vocal functioning subscale (items 3, 4, 5, 9, 10, 11, 15, 16, 18, 19, 20, 21, 24, 29), a 12-item social participation subscale (items 1, 2, 6, 7, 8, 12, 13, 14, 17, 23, 25, 30), and 4 items that loaded onto both (items 22, 26, 27, 28).⁴

†Bultynck et al.'s three-factor model consisted of an 11-item anxiety and avoidance subscale (items 2, 7, 8, 12, 13, 16, 17, 23, 25, 26, 30), an 8-item vocal identity subscale (items 3, 4, 6, 10, 19, 20, 24, 28), and an 11-item vocal functioning subscale (items 1, 5, 9, 11, 14, 15, 18, 21, 22, 27, 29).⁸

not unique dimensions that function as distinct traits, and it is reasonable to interpret the TWVQ as unidimensional. However, to investigate if the item groups identified in the studies could aid score interpretation as descriptive categories rather than separate factors, they were overlaid onto the Wright Map organized by ascending first thresholds (Fig. 4). When using Dacakis et al.'s groupings,⁴ those items labeled as “vocal functioning” tended to correspond to items with lower first thresholds than those labeled “social participation.” When using Bultynck et al.'s groupings,⁸ those items labeled as “vocal

identity” tended to have lower first thresholds than items labeled as “vocal function” and “anxiety-avoidance.”

Relationship to VHI-10 (Model 3)

Latent regression revealed that the VHI-10 was a significant predictor of TWVQ scores (coefficient = 0.10, SE = 0.014, 95% confidence interval: 0.07–0.13); for every 10-point increase in VHI-10 score, the TWVQ score is predicted to increase by 1 logit, on average. When overlaid onto the revised Wright Map (Fig. 3), VHI-10 overlapped with the upper range of the TWVQ (−1.74 to +2.34 logits, scaled score 110–192) but not the lower range, suggesting the TWVQ can detect lower levels of voice-related life impact in trans women compared to VHI-10.

DISCUSSION

The results of this study revealed several insights and improvements to the TWVQ. First, using the Rasch PCM, responses were transformed from ordinal items into an interval logit scale and a subsequent scaled score representing voice-related life impact for trans women. Now, unlike raw scores, differences between scaled scores are equivalent throughout the scale. This calibration can be useful to compare pre- and post-treatment score changes between patients starting at different pre-treatment levels.

Second, the Wright Map for the original scoring system was challenging to interpret given the extensive overlap of the second threshold with the others, suggesting the response categories “sometimes” and “often” were not meaningfully different. This finding is consistent with prior reports of response patterns, which revealed that none of

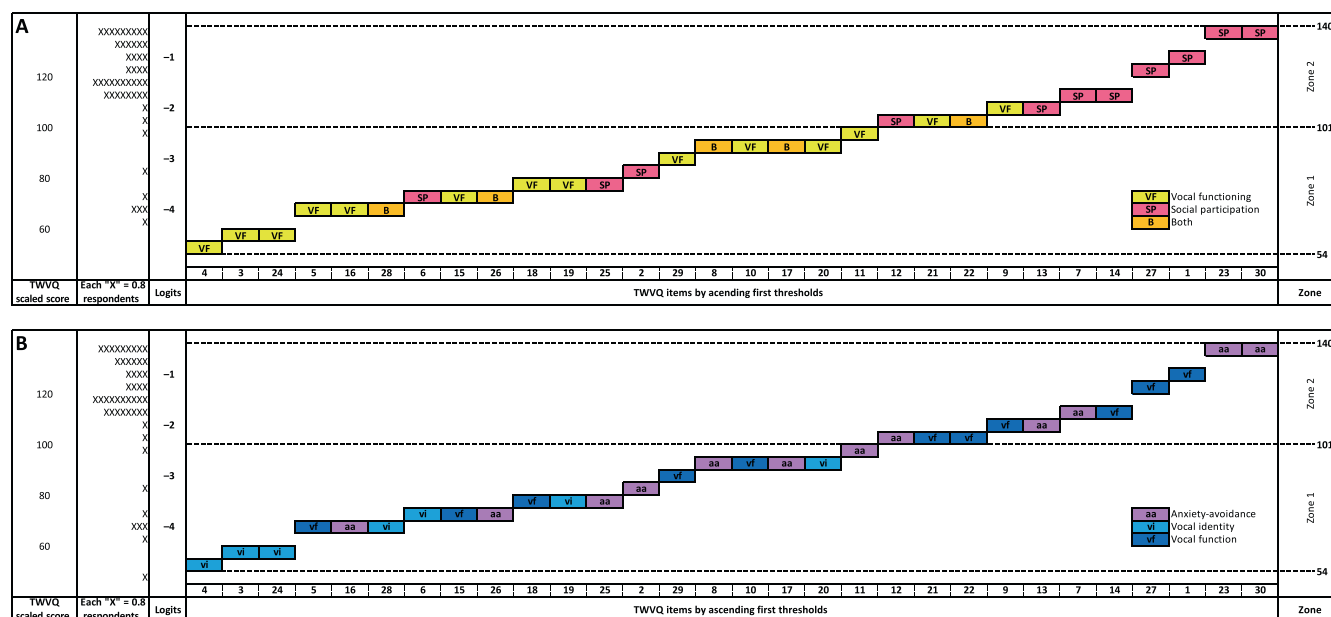


Fig. 4. Partial Wright Maps depicting the relationship between item groups from Dacakis et al.⁴ (A) and Bultynck et al.⁸ (B) and item first threshold locations based on updated scoring. Each “X” represents a group of 0.8 respondents. Thick dashed lines depict interpretive zones 1 (>54 to ≤101) and 2 (>101 to ≤140) that correspond with the locations of item first thresholds. TWVQ = Trans Woman Voice Questionnaire. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

the items had a modal response of 3 (often).¹ The updated scoring collapsed “sometimes” and “often” into one score category and improved evidence of internal structure through threshold banding. Scoring procedures rather than the instrument form were changed because altering format can alter responses.^{28,29} Future work can examine if simplifying the rating form can reduce respondent burden and retain validity.

The updated scoring also aids interpretation because there is now empirical evidence for four scaled score zones referenced to the items: ≤ 54 (Zone 0), >54 to ≤ 101 (Zone 1), >101 to ≤ 140 (Zone 2), and >140 (Zone 3). Each of these zones corresponds to a pattern of increasing response levels; therefore, it may be useful to interpret these scoring ranges as minimal, low, moderate, and high voice-related life impact for trans women seeking gender-affirming voice care. In addition, the item categories identified through factor analyses, particularly Dacakis et al.’s work,⁴ help add an additional layer of meaning to the zones: respondents scoring in Zone 1 are primarily endorsing “vocal functioning” items, while those scoring in Zones 2 are endorsing more and more “social participation” items. Although further research is needed to understand if these regions are meaningful for either decision-making or assessing treatment response, it is important to recognize these patterns were only visible when evaluating the Wright Map, underscoring its value to instrument development.

Third, item fit statistics highlighted an opportunity to examine the items themselves for improvements. Although three items were outside of the conventional range, item 1 (people have difficulty hearing me in a noisy room) was the most problematic. An elevated WMNSQ value indicates the responses to the item are not well predicted by the model.^{15,17} Therefore, there may be another sources of variance within the data, such as inconsistencies in how individuals are interpreting the item or that the item does not fully capture patient perceptions. However, a decision to remove an item should not be based on the WMSQ values alone but rather on the contribution of the item to the overall construct. Further investigation should be conducted, such as repeated sampling or respondent cognitive interviews, to determine if this and other items should be kept, revised, or removed.^{15,17}

Fourth, multidimensional analysis (analogous to confirmatory factor analysis) demonstrated strong latent correlations between previously proposed subscales of the TWVQ,^{4,8} supporting a unidimensional interpretation. Though subscales can be useful, there is a practical advantage to maintaining a unidimensional lens: a single Wright Map. Viewing all items on one Wright Map affords us multiple insights into the structure and function of the instrument, such as those presented in this article.

Finally, this study provides new insight into the differences between the TWVQ and the VHI-10. Similar to other studies, there was a significant relationship between the TWVQ and the VHI-10.⁷ However, latent regression revealed that the VHI-10 primarily covered the second threshold across the items. Therefore, the benefit of the TWVQ is that it can stratify trans women who are experiencing lower levels of voice impact that would

otherwise score similarly on the VHI-10. Clinicians can now use this knowledge to make informed decisions regarding the trade-offs of administering one or both instruments in their practices.

Recommendations for Use

Given the findings from this study, we recommend that the TWVQ be completed in its current form with the original four answer choices. However, scoring of the instrument should be altered, coding “never or rarely,” “sometimes,” “often,” and “usually or always” as 0, 1, 1, and 2, respectively. The raw score can then be converted to logits and/or the scaled score using the conversion table (Table S1). Importantly, the conversion table can only be used for patients who have completed all items; questionnaires with missing data must be scored using IRT software. This new scoring seeks to enhance the utility of the TWVQ for clinical care and guiding patient-centered discussions, not for gatekeeping purposes of who should or should not receive care.

Limitations

The sample size in this study is smaller than what some have advocated for IRT models.³⁰ However, the Rasch PCM does not estimate as many parameters as other models (e.g., 2-parameter logistic) and can therefore converge with smaller samples.²² In addition, the trans women population is relatively small compared to other populations of interest; as a result, we believe it is reasonable to draw conclusions about a treatment-seeking trans women population. Nevertheless, whether these findings extend to the broader trans women community remains to be seen, especially given our sample did not contain many respondents with low scores.

Toward a Theory for Score Interpretation

Ultimately, this study has generated more questions than answers. Despite improving score meaning, it remains a challenge to extend interpretation beyond minimal, low, moderate, and high score groups. That is, what does it mean for the patient to experience a voice-related life impact at different levels? The relationship with the item categories provides some insight into how interpretation can improve with an overarching theory of voice-related life impact for trans women, where those individuals with lower scores may primarily experience issues related to vocal function or identity and those scoring higher may experience additional anxiety and impacts related to social participation. More in-depth work is required to fully develop a theory to support this observation, but once a strong theory of voice-related life impact for trans women is proposed, construct modeling can be used to develop more useful instruments. In this process, the construct to be measured is laid out *a priori* along a continuum called the construct map, in qualitatively distinct levels.^{15–17,23,24} Items are then developed to reflect these theoretical levels, and empirical findings through IRT can then be used to understand how well the items fit the construct. Applying IRT in conjunction with a clear construct as a

theory for measurement can advance instrument development and measurement for both research and clinical care.

CONCLUSION

Empirical data from Rasch analyses led to updated scoring and interpretive zones for the TWVQ that will advance the clinical and research utility of the instrument. The IRT methods and interpretation-focused Wright Map analyses used in this study can be used to enhance score meaning for other patient measures developed without theoretically-derived construct levels, improving patient care and outcomes assessment in laryngology, speech-language pathology, and beyond.

BIBLIOGRAPHY

- Dacakis G, Davies S, Oates JM, Douglas JM, Johnston JR. Development and preliminary evaluation of the transsexual voice questionnaire for male-to-female transsexuals. *J Voice*. 2013;27(3):312-320. <https://doi.org/10.1016/j.jvoice.2012.11.005>.
- T'Sjoen G, Moerman M, Van Borsel J, et al. Impact of voice in transsexuals. *Int J Transgend*. 2006;9(1):1-7. https://doi.org/10.1300/J485v09n01_01.
- Dacakis G, Oates JM, Douglas JM. Exploring the validity of the Transsexual Voice Questionnaire (male-to-female): do TVQMtF scores differentiate between MtF women who have had gender reassignment surgery and those who have not? *Int J Transgend*. 2016;17(3-4):124-130. <https://doi.org/10.1080/15532739.2016.1222922>.
- Dacakis G, Oates JM, Douglas JM. Further evidence of the construct validity of the transsexual voice questionnaire (TVQMtF) using principal components analysis. *J Voice*. 2017;31(2):142-148. <https://doi.org/10.1016/j.jvoice.2016.07.001>.
- Dacakis G, Oates J, Douglas J. Associations between the Transsexual Voice Questionnaire (TVQMtF) and self-report of voice femininity and acoustic voice measures. *Int J Lang Commun Disord*. 2017;52(6):831-838. <https://doi.org/10.1111/1460-6984.12319>.
- Brown SK, Chang J, Hu S, et al. Addition of Wendler glottoplasty to voice therapy improves trans female voice outcomes. *Laryngoscope*. 2021; 131(7):1588-1593. <https://doi.org/10.1002/lary.29050>.
- Young VN, Yousef A, Zhao NW, Schneider SL. Voice and stroboscopic characteristics in transgender patients seeking gender-affirming voice care. *Laryngoscope*. 2020;131:1071-1077. <https://doi.org/10.1002/lary.28932>.
- Bultynck C, Pas C, Defreyne J, Cosyns M, T'Sjoen G. Organizing the voice questionnaire for transgender persons. *Int J Transgend Health*. 2020; 21(1):89-97. <https://doi.org/10.1080/15532739.2019.1605555>.
- Rosen CA, Lee AS, Osborne J, Zullo T, Murry T. Development and validation of the Voice Handicap Index-10. *Laryngoscope*. 2004;114(9):1549-1556. <https://doi.org/10.1097/00005537-200409000-00009>.
- Magno C. Demonstrating the difference between classical test theory and item response theory using derived test data. *Int J Educ Psychol Assess*. 2009;1(1):1-11.
- Rusch T, Lowry PB, Mair P, Treiblmaier H. Breaking free from the limitations of classical test theory: developing and measuring information systems scales using item response theory. *Inf Manage*. 2017;54(2): 189-2013. <https://doi.org/10.1016/j.im.2016.06.005>.
- Embretson SE. The new rules of measurement. *Psychol Assess*. 1996;8(4): 341-349. <https://doi.org/10.1037/1040-3590.8.4.341>.
- Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. *Patient*. 2014;7(1):23-35. <https://doi.org/10.1007/s40271-013-0041-0>.
- Columbia Public Health. Item response theory. Accessed June 19, 2022. <https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory#:~:text=Courses-,Overview,outcomes%2C%20responses%20or%20performance>
- Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Educ Res*. 2006;21(suppl 1):i4-i18. <https://doi.org/10.1093/her/cyl108>.
- Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health Educ Res*. 2006;21(suppl 1):i19-i32. <https://doi.org/10.1093/her/cyl053>.
- Wilson M. *Constructing Measures: An Item Response Modeling Approach*. Routledge; 2004.
- Messick S. Foundations of validity: meaning and consequences in psychological assessment. *ETS Res Rep Ser*. 1993;1993(2):i-18. <https://doi.org/10.1002/j.2333-8504.1993.tb01562.x>.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014.
- Adams R, Wu M, Cloney D, Wilson M. *ACER ConQuest: Generalised Item Response Modelling Software*. Australian Council for Educational Research; 2020.
- Masters GN. A rasch model for partial credit scoring. *Psychometrika*. 1982; 47(2):149-174. <https://doi.org/10.1007/BF02296272>.
- Masters GN, Wright BD. The partial credit model. In: van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory*. Springer; 1997:101-121. https://doi.org/10.1007/978-1-4757-2691-6_6.
- Brondfield S, Blum A, Lee K, Linn M, O'Sullivan PS. The cognitive load of inpatient consults: development of the consult cognitive load instrument and initial validity evidence. *Acad Med*. 2021;96:1732-1741. <https://doi.org/10.1097/ACM.00000000000004178>.
- Blum AM, Mason JM, Kim J, Pearson PD. Modeling question-answer relations: the development of the integrative inferential reasoning comic assessment. *Read Writ*. 2020;33(8):1971-2000. <https://doi.org/10.1007/s11145-020-10026-4>.
- Adams RJ, Wilson M, Wang W. The multidimensional random coefficients multinomial logit model. *Appl Psychol Measur*. 1997;21(1):1-23. <https://doi.org/10.1177/0146621697211001>.
- Briggs DC, Wilson M. An introduction to multidimensional measurement using Rasch models. *J Appl Meas*. 2003;4(1):87-100.
- de Boeck P, Wilson M. A framework for item response models. In: de Boeck P, Wilson M, eds. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer; 2004:3-41. https://doi.org/10.1007/978-1-4757-3990-9_1.
- Stewart AL, Thrasher AD, Goldberg J, Shea JA. A framework for understanding modifications to measures for diverse populations. *J Aging Health*. 2012;24(6):992-1017. <https://doi.org/10.1177/0898264312440321>.
- Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report. *Value Health*. 2009;12(8):1075-1083. <https://doi.org/10.1111/j.1524-4733.2009.00603.x>.
- Baylor C, Hula W, Donovan NJ, Doyle PJ, Kendall D, Yorkston K. An introduction to item response theory and Rasch models for speech-language pathologists. *Am J Speech Lang Pathol*. 2011;20(3):243-259. [https://doi.org/10.1044/1058-0360\(2011/10-0079\)](https://doi.org/10.1044/1058-0360(2011/10-0079)).