VALUE IN HEALTH 18 (2015) 25–34

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Patient-Reported Outcomes

Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples

Jennifer Petrillo, PhD[1,], Stefan J. Cano, PhD[2], Lori D. McLeod, PhD[3], Cheryl D. Coon, PhD[3]*

[1]*Novartis AG, Basel, Switzerland;* [2]*Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK;* [3]*RTI Health Solutions, Research Triangle Park, NC, USA*

ABSTRACT

Objective: To provide comparisons and a worked example of item- and scale-level evaluations based on three psychometric methods used in patient-reported outcome development—classical test theory (CTT), item response theory (IRT), and Rasch measurement theory (RMT)—in an analysis of the National Eye Institute Visual Functioning Questionnaire (VFQ-25). **Methods:** Baseline VFQ-25 data from 240 participants with diabetic macular edema from a randomized, double-masked, multicenter clinical trial were used to evaluate the VFQ at the total score level. CTT, RMT, and IRT evaluations were conducted, and results were assessed in a head-to-head comparison. **Results:** Results were similar across the three methods, with IRT and RMT providing more detailed diagnostic information on how to improve the scale. CTT led to the identification of two problematic items that threaten the validity of the overall scale score, sets of redundant items, and skewed response categories. IRT and RMT additionally identified poor fit for one item, many locally dependent items, poor targeting, and disordering of over half the response categories. **Conclusions:** Selection of a psychometric approach depends on many factors. Researchers should justify their evaluation method and consider the intended audience. If the instrument is being developed for descriptive purposes and on a restricted budget, a cursory examination of the CTT-based psychometric properties may be all that is possible. In a high-stakes situation, such as the development of a patient-reported outcome instrument for consideration in pharmaceutical labeling, however, a thorough psychometric evaluation including IRT or RMT should be considered, with final item-level decisions made on the basis of both quantitative and qualitative results.

Keywords: patient-reported outcome, psychometrics, item response theory, classical test theory, Rasch measurement theory.

Introduction

The patient perspective is increasingly critical in clinical decision making as a means to communicate treatment benefit [1]. To fulfill this role, patient-reported outcome (PRO) instruments have been used to quantify how patients feel and function [2]. More recently, to support drug development, PROs have been increasingly included as key primary or secondary end points in clinical studies [3]. Given the growing prominence of PRO instruments, it is essential that they are scientifically robust and fit for purpose [4]. Although qualitative research drives content development for PRO instruments, the role of quantitative psychometric methods is to test measurement performance. There are three main psychometric paradigms: classical test theory (CTT) [5], item response theory (IRT) [6], and Rasch measurement theory (RMT) [7].

In brief, CTT can be traced back to Spearman at the turn of the 20th century who introduced the decomposition of an observed score into a true score and an error and estimated the reliability of observed scores [8]. The premise is that items can be summed (without weighting or standardization) to produce a total score [6]. IRT can be traced back to Thurstone's law of comparative judgment from the 1920s and was fully expounded in the work of Lord [9]. Its foundations lie in the use of stochastic models to derive statistical estimation of parameters that represent the locations of persons and items on a latent continuum [10]. RMT was born out of the work of Rasch in the middle of the 20th century. He developed the simple logistic model (now known as the Rasch model), and through applications in education and psychology, he argued that he was able to demonstrate, mathematically, that his approach met stringent criteria for measurement used in the physical sciences [7]. The key difference between the approaches is that CTT and IRT typically describe a set of data, whereas RMT aims to obtain data that fit the model. A fuller description of the three paradigms can be found

Conflicts of interest: Dr. Petrillo is currently employed at Biogen Idec, Cambridge, MA, USA. Dr. Coon is currently employed at Adelphi Values, Boston, MA, USA. Dr. Cano is currently co-founder at Modus Outcomes, Newton, MA, USA.

* *Address correspondence to:* Lori McLeod, RTI Health Solutions, 3040 Cornwallis Road, Research Triangle Park, NC 27709, USA.
E-mail: lmcleod@rti.org.

1098-3015$36.00 – see front matter © 2015 Published by Elsevier Inc. on behalf of International Society for Pharmacoeconomics and Outcomes Research (ISPOR).
http://dx.doi.org/10.1016/j.jval.2014.10.005

elsewhere (CTT [6,11–13], RMT [14–19], IRT [20–24]). In addition, Table 1 provides additional background and comparisons across the approaches.

Direct comparisons of these psychometric methods in the PRO instrument literature are rare (and in the wider clinical outcome assessment literature, of which PRO instruments are a subtype, almost nonexistent). In part, this is because these paradigms encompass different methodologies, produce different information, and apply different criteria for success and failure.

Each approach has its supporters, and the traditional psychometric approach (CTT) remains the dominant paradigm. However, with IRT and RMT increasingly gaining favor in clinical outcomes assessment, understanding the value of each of these different psychometric approaches is essential. In this study, we compare CTT, IRT, and RMT in an analysis of the National Eye Institute Visual Functioning Questionnaire (VFQ-25) [31], a widely used PRO instrument purporting to measure visual functioning in patients with vision impairment. The focus of this research was to describe typical analyses specific to each of the three paradigms and then to apply these analyses to a common data set, adopting the criteria for each paradigm separately rather than forcing the same criteria across the set. This was done to maintain the integrity of each method before comparing results for item- and scale-level evaluations. This research was not designed to evaluate the dimensionality of the example instrument, which would be better served in further research.

Methods

Sample

Baseline VFQ-25 data from 240 participants were made available from the clinical trial dataset study, a randomized, double-masked, multicenter, laser-controlled phase 3 study of an injectable in patients with visual impairment due to diabetic macular edema [32,33]. Patients qualified for the study if their best-corrected visual acuity ranged from 78 to 39 letters (i.e., Snellen-equivalent 20/32 to 20/160).

PRO Measure

The VFQ-25 includes 25 items plus a single-item general health rating. Items are grouped in the following proposed 12 scales: general health (1 item), near vision (3 items), ocular pain (3 items), distance vision (3 items), vision-specific social functioning (2 items), vision-specific mental health (4 items), vision-specific role functioning (2 items), expectations for visual function (3 items), dependency due to vision (3 items), driving (3 items), peripheral vision (1 item), and color vision (1 item). Items are scored by patients on five- or six-point response options. A total score is calculated by summing responses to items in each of the scales (score range 0–100), with higher scores indicating better functioning.

CTT methods were used when the VFQ-25 was originally developed [31]; however, recent studies have explored its properties using IRT and RMT [29,34–36], which has led to mixed results. To date, the VFQ-25 has not been subjected to a head-to-head comparison of CTT, IRT, and RMT. For the purpose of this worked example, the original scoring and proposed scales of the VFQ-25 were ignored, and a single underlying construct of visual functioning was assumed.

Data Analysis

Three psychometricians (coauthors: L.D.M., C.D.C., and S.J.C.) independently conducted CTT, IRT, and RMT analyses, respectively, on the VFQ-25 items in the clinical trial data set. Although each method evaluates its own unique set of instrument properties, a subset of comparable properties was selected to facilitate a head-to-head comparison.

Classical test theory

CTT analyses were conducted and graphics produced in Microsoft Excel 2007. For the CTT evaluation of the VFQ-25, the following analyses were carried out:

1. *Data quality and scaling evaluation*: The amount of missing item-level data was used to evaluate data quality (no more than 5% missing was considered acceptable). Item-level response descriptive statistics were used to evaluate whether the item response categories were being used appropriately (each response category should have been selected by a subset of patients) without evidence for floor or ceiling effects (at least 5% but not more than 40% selecting the extreme categories was considered acceptable). Pattern of item mean scores over time was measured (skewness of $\leq |2|$ was considered acceptable [37]). A trend for greater amounts of missing data at the end of the instrument was a flag for respondent fatigue. A trend for greater amounts of missing data by content was a flag for potential content validity issues.
2. *Scaling assumptions*: Inter-item correlations and item-to-total correlations were used to gauge the strength of the relationships among the items and the appropriateness of scoring them together on one scale, as well as evidence for local dependencies. Inter-item correlations of 0.8 or less and item-to-total correlations of 0.3 or more were considered acceptable.
3. *Reliability*: Cronbach's alpha [30] internal consistency reliability was used to assess the degree to which items are related. A value of $\alpha \geq 0.70$ was considered acceptable.

Item response theory

To evaluate the VFQ-25 using an IRT model, the item parameters were calibrated and associated statistics and graphics were produced using IRTPRO (version 2.1) [38]. Samejima's [39] graded response model was selected, which assumes variable slope parameters across the items on the scale. For the IRT evaluation of the VFQ-25, the following analyses were carried out:

1. Item characteristic curves graphically show the probability of an item response across the range of the scale and reveal weak items (i.e., low slopes) or overlapping response categories.
2. *Item fit*: S-X^2 reflects the differences between observed and expected response proportions for each test score value. Significant values indicate items with potential misfit [40,41].
3. *Local dependence*: X^2 examines the bivariate fit to identify evidence of potentially redundant items. Values larger than 10 indicate likely local dependence, whereas values between 5 and 10 may suggest local dependence or may be a result of sparseness in the frequency table [42].
4. *Reliability*: Item and test information functions graphically reflect how reliably the individual items and the test as a whole estimate the construct over the entire scale range. Values can be converted into an estimate of reliability (i.e., reliability = 1 − [1 / information]) so that the common rule of thumb of 0.70 to 0.90 for interpreting reliability values corresponds to information of 3.3 to 10 [43].

Rasch measurement theory

RMT methods were implemented using RUMM2030 software. The analyses conducted are described below:

1. *Targeting*: Scale-to-sample targeting concerns the match between the range of health impact due to vision problems

Table 1 – CTT, RMT, and IRT: Comparison of evaluations.

Psychometric property	CTT Evaluation [25–27]	IRT Evaluation [6,29]	RMT Evaluation [16,17,28]*
Acceptability	The percentage of missing data for each item and the percentage of people for whom a PRO instrument score can be computed	There are no formal RMT analyses for this property of a PRO instrument	There are no formal IRT analyses for this property of a PRO instrument
Targeting of the items	PRO instrument scores should span the entire range; floor (proportion of the sample at the maximum score) and ceiling (proportion of the sample at the minimum score) effects should be low	The PRO items should provide information across the full range of the population for which it is intended	The relative distributions of item locations and person estimates (statistical indicators) are examined statistically and graphically
Scaling assumptions	Summing item scores is considered legitimate, when the items: • Are approximately parallel (i.e., they measure at the same point) • Contribute similarly to the variation of the total score (i.e., similar variances); otherwise, these should be standardized • Measure a common underlying construct • Contain a similar proportion of information concerning the construct being measured	There are no formal IRT analyses for this property of a PRO instrument	There are no formal RMT analyses for this property of a PRO instrument
Suitability of the response options	There are no formal traditional analyses for this property of a PRO instrument, although the patterns of item endorsement frequencies can be examined	Each response option should provide information within the range of the population for which it is intended. Each response option should be distinct and should have a range along the scale within which it is the most likely response choice	The examination of category probability curves show the ordering of the thresholds for each item. A threshold marks the location on the latent continuum where two adjacent response categories are equally likely. The ordering of the thresholds should reflect the intended order of the categories
Validity	The validity of the scale is evaluated using inter-item correlations and item-to-total correlations to gauge the strength of the relationships among the items and the appropriateness of scoring them together on one scale as well as evidence for local dependencies	Broad internal validity indicators.	Broad internal validity indicators.
		Fit residuals (statistical) summarize the difference between observed and expected responses to an item across all people (item–person interaction). Item characteristic curves display graphically the expected responses for each item across the continuum (the curve).	Fit residuals (statistical) summarize the difference between observed and expected responses to an item across all people (item–person interaction). Chi-square values (statistical) summarize the difference between observed and expected responses to an item for groups (known as class intervals) of people with relatively similar levels of ability (item–trait interaction). Item characteristic curves display graphically the expected responses for each item across the continuum (the curve), and the mean observed scores for each group of person scores (class

continued on next page

Table 1 – *continued*

Psychometric property	CTT Evaluation [25–27]	IRT Evaluation [6,29]	RMT Evaluation [16,17,28]*
			intervals) can be plotted against the item characteristic curves.
		Local independence: IRT analyses also consider the item scoring bias resulting from similar items being included in the same instrument.	Local independence: RMT analyses also consider item scoring bias, which is the extent to which items are locally independent, i.e., individual items are not biased by each other
Reliability	Commonly assessed using Cronbach's alpha coefficient [30] and item internal consistency indicators, including item-total correlations	Assessed using the information curve, which is analogous to Cronbach's alpha being calculated separately at each score along the range of the scale. This reflects that an instrument's reliability may change depending on the level of the underlying condition being measured	Examined using the Person Separation Index, which is analogous to Cronbach's alpha

CTT, classical test theory; IRT, item response theory; PRO, patient-reported outcome; RMT, Rasch measurement theory.
* Although the general tenet around issues such as fit, dependency, and reliability are effectively consistent across Rasch-based software programs, the broad descriptions here are based on analyses and outputs generated through RUMM 2030 and cannot be considered as exactly the same as other programs.

measured by the VFQ-25 items and the range of health impact due to vision problems in the sample of patients.

2. *Ordering of item thresholds*: Each of the items of the VFQ-25 has multiple response categories that reflect an ordered continuum of health impact due to vision problems (e.g., 1, 2, 3, 4, 5…). Although this ordering may appear clinically sensible at the item level, it must work when the items are combined to form a set. Item fit validity analysis tests this statistically and graphically by threshold locations and plots. As such, we would expect the threshold values between adjacent pairs of response options to be ordered by magnitude (less to more). This is visible in graphical plots, in which the highest areas of the probability distributions of each response category should not be below either adjacent category plots.
3. *Item fit validity*: The items of the VFQ-25 must work together (fit) as a conformable set both clinically and statistically. Otherwise, it is inappropriate to sum item responses to reach a total score and consider the total score as a measure of health impact due to vision problems. When items do not work together in this way (misfit), the validity of a scale is questioned. Three main indicators can be examined to assess this: 1) fit residuals (item–person interaction), 2) chi-square values (item–trait interaction), and 3) item characteristic curves. There are no absolute criteria for interpreting fit statistics. Fit residual should fall between −2.5 and +2.5 with associated nonsignificant chi-square values (significance interpreted after Bonferroni adjustment) [28]. It is more meaningful to interpret these statistics together and in the context of their clinical usefulness as an item set.
4. *Item dependency*: The response to one VFQ-25 item should not directly influence the response to another. Item dependency determines this effect by examining the residual correlations ($r > 0.30$ indicates potential dependency) [28].
5. *Reliability*: Examination of person measurements (estimates) and the Person Separation Index (a reliability statistic comparable to Cronbach's alpha that quantifies the error associated with the measurements of people in this sample). Higher Person Separation Index values indicate better reliability (>0.70 indicates adequate reliability) [44].

Results

Classical Test Theory

Table 2 provides the CTT item-level descriptive statistics and evaluation results. Items were reverse scored so that higher scores are indicative of better functioning and performance. The item response frequency distribution was also reviewed, first sorted by the order in which the items were administered and then by mean item response (Fig. 1).

1. Data quality and scaling
   - A few items were found to have no responses in the highest category. For example, no participants selected the first category for item 2 (general vision), which would indicate blindness. Upon review, this is not a concern, because it is due to the eyesight exclusion criteria for the trial.
   - Multiple items were flagged for evidence of either floor or ceiling effects. Overall, there is little use of the higher categories, indicating that the item set may be more appropriate for a population with more severe impairment than our pretreatment sample had.
   - Most items were negatively skewed; the item response distributions were mainly in the upper end of the scale. Mean responses were "3" or above for each item except for item 3 (worry). Item 12 (matching clothes) and item 13 (visiting others) were flagged for extreme skewness.
   - Items 15c to 16a (driving) had a higher level of missing data likely due to participants' not currently driving.
   - About 80% of the participants responded as having no difficulty picking out or matching clothes because of their vision (item 12). This question may not be as effective for discriminating because of its social or fashion perspective.

Table 2 – CTT: VFQ-25 Item descriptive and correlation results (reverse scored).

Item	N	Miss	Mean ± SD	Skew	Minimum	Maximum	Inter-item corrs	Item-total corrs	Flags*
2—General vision (1–6)	239	1	4.02 ± 0.72	–0.24	2	6		0.58	
3—Worry (1–5)	240	0	2.58 ± 1.15	0.28	1	5		0.45	
4—Amount of pain (1–5)	240	0	4.33 ± 0.83	–1.03	1	5	18 r's < 0.3	0.25	Floor, r total
5—Reading normal newsprint (1–5)	238	2	3.13 ± 1.25	–0.19	1	5		0.56	
6—Seeing well up close (1–5)	240	0	3.49 ± 1.14	–0.31	1	5		0.67	
7—Finding objects on crowded shelf (1–5)	239	1	4.10 ± 0.98	–0.87	1	5		0.69	Floor
8—Reading street signs (1–5)	238	2	3.96 ± 0.98	–0.62	1	5		0.61	
9—Going down stairs at night (1–5)	238	2	3.84 ± 1.06	–0.51	1	5		0.62	
10—Seeing objects off to side (1–5)	239	1	4.19 ± 0.99	–1.07	1	5		0.71	Floor
11—Seeing how people react (1–5)	240	0	4.39 ± 0.88	–1.46	1	5		0.62	Floor
12—Difficulty matching clothes (1–5)	234	6	4.68 ± 0.71	–2.23	2	5		0.65	Floor, skew
13—Visiting others (1–5)	238	2	4.54 ± 0.86	–2.10	1	5		0.55	Floor, skew
14—Going out to movies/plays (1–5)	208	32	4.17 ± 1.09	–1.29	1	5		0.75	Miss, floor
15c—Daylight driving (1–4)	158	82	4.09 ± 1.56	–1.38	1	5	$r = 0.88$ (item 16) $r = 0.90$ (item 16a)	0.57	Miss, r item
16—Nighttime driving (1–5)	157	83	3.18 ± 1.47	–0.36	1	5	$r = 0.88$ (items 15c and 16a)	0.61	Miss, r item
16a—Difficult conditions driving (1–5)	158	82	3.44 ± 1.49	–0.58	1	5	$r = 0.88$ (item 15c) $r = 0.90$ (item 16)	0.63	Miss, r item
17—Accomplish less (1–5)	239	1	3.53 ± 1.27	–0.41	1	5		0.65	
18—Limited in endurance (1–5)	238	2	3.89 ± 1.19	–0.77	1	5		0.71	Floor
19—Amount of time without pain (1–5)	239	1	4.44 ± 0.93	–1.78	1	5		0.34	Floor
20—Stay home most of time (1–5)	238	2	4.30 ± 1.18	–1.58	1	5		0.72	Floor
21—Frustrated (1–5)	237	3	3.72 ± 1.32	–0.64	1	5		0.68	
22—No control (1–5)	238	2	3.83 ± 1.40	–0.79	1	5		0.72	
23—Rely too much on others' word (1–5)	238	2	4.29 ± 1.18	–1.55	1	5	Correlated 0.88 with item 24	0.63	Floor, r item
24—Need much help from others (1–5)	238	2	4.24 ± 1.23	–1.45	1	5	Correlated 0.88 with item 23	0.64	Floor, r item
25—Embarrassment (1–5)	238	2	4.38 ± 1.09	–1.63	1	5		0.61	Floor

Corrs, correlation; CTT, classical test theory; VFQ-25, National Eye Institute Visual Functioning Questionnaire.

* miss, flagged for missing; floor, flagged for floor or ceiling; skew, flagged for skewness; r item, flagged for >0.80 inter-item correlations; r total, flagged for <0.30 correlation with total score.

- Very little pain was reported, with most responses indicating mild or no pain (items 4 and 19).
- There was no evidence to support a fatigue effect due to the length of the questionnaire. Instead, we found evidence for the response patterns to group by content. For example, the response distributions are indicative of the "social issues" (e.g., item 12 and item 13) being the least impacted by vision.

2. Scaling assumptions
   - The three items involving driving (items 15c–16a) were highly related. One potential solution for these items is to score these

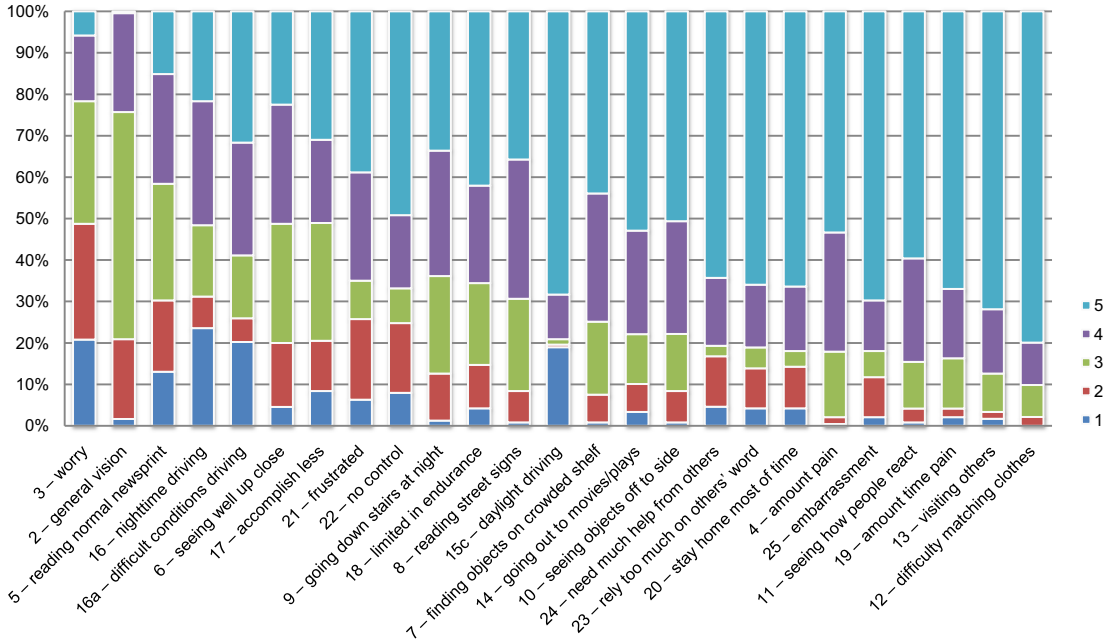

Fig. 1 – CTT: VFQ-25 Item response frequencies by mean item response. CTT, classical test theory; VFQ-25, National Eye Institute Visual Functioning Questionnaire.

items together and report one score for the composite (or triplet) formed by combining these three items.
- Item 23 (rely on others' word) and item 24 (help from others) both measured social interactions and were flagged for being correlated (>0.80).
- Item 4 (amount of pain) and item 19 (amount of time without pain), although they are related to each other, are the least related to the other items.

3. Reliability
   - Alpha is quite high for these 25 items at 0.94. The item-total correlation results support our earlier findings, with the addition that the pain items are the least related to the total at $r = 0.25$ for item 4 and $r = 0.34$ for item 19 (Table 2). To investigate possible scale revisions, alpha was computed for a 22-item version, removing the pain items (item 4 and item 19) and the matching clothes item (item 12). Alpha remained at 0.92, with all items correlated at least 0.4 with the total.

Recommendations to consider on the basis of the CTT results include removing item 12 related to matching clothes as well as scoring the driving items (items 15c–16a) together and reporting one score for these three within the final scale. Alternatively, one of the driving items could be selected to represent the concept of driving impact rather than a composite based on the full set of three. In addition, given the low correlations between the pain items (item 4 and item 19) and the rest of the item set, consideration should be given to reporting the pain items as a separate score rather than combining them into the total score. Finally, depending on decisions about content and the final target population, the items related to social functioning should be considered for additional omission because of their skewness and floor effects (items 13, 14, 23, and 24).

Item Response Theory

The item characteristic curves, information functions, and S-X^2 P values for two items are presented in Figure 2 to provide visual examples of a good item (item 3—worry) and a problematic item (item 20—stay home). The test information function and standard error for the original 25 items are presented in the left side of Figure 3. Results across all 25 items are as follows:

1. *Slopes*: Item 4 (amount of pain) and item 19 (amount of time without pain) had very weak slopes, indicating poor information.
2. *Thresholds*: Most of the items had great overlap among many of the response categories (e.g., item 20—stay home), resulting in those categories offering little in terms of placing people on the scale.
3. S-X^2: A total of 18 of the 25 items (e.g., item 20—stay home) indicated some degree of misfit ($P < 0.05$).
4. *Local dependence* X^2: Four groups of items had questionable local dependence values ($X^2 \geq 5$) and similar item content (i.e., pain [items 4 and 19], seeing up close [items 5 and 6], driving [items 15c, 16, and 16a], and relying on others [items 23 and 24]).
5. *Test information*: The VFQ-25 provides adequate measurement (reliability ≥ 0.70, i.e., $I \geq 3.3$) for all but the very highest part of theta (i.e., best visual functioning).

Based on the data at hand, the IRT results suggest a number of item and instrument revisions. Response category revisions are suggested for items 12, 13, and 17 through 25, which appeared to be used dichotomously. Furthermore, including only the one driving item focused on the ability to drive at night would be the most informative and least redundant way to include the concept of driving. Item 23 (rely on others' word) is recommended for deletion because it covers a concept already measured in more informative item 24 (need help from others). The two items concerning pain or discomfort, 4 and 19, are also recommended for deletion because they provided inconsistent information because they measure a symptom and this scale is specific to functioning.

Rasch Measurement Theory

1. *Targeting*: Distributions of item thresholds and person estimates were relatively well matched, but there was significant item bunching and some gaps at the highest end of the

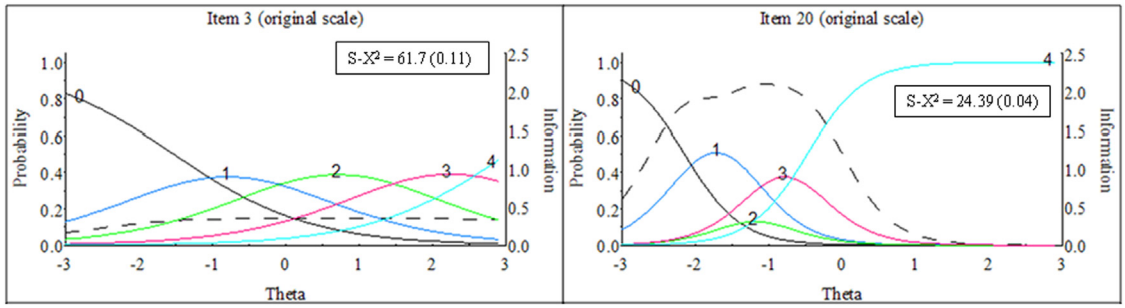

Fig. 2 – IRT: VFQ-25 Item characteristic curves and information functions from the graded-response model. The colored curves are item characteristic curves, each corresponding to a different response category, and the dashed lines are item information functions. Item 3 (worry) is an example of a good item, whereas item 20 (stay home) is an example of a problematic item. IRT, item response theory; VFQ-25, National Eye Institute Visual Functioning Questionnaire.

continuum. This indicates a significant ceiling effect. The peak of the information plot was just above 0 logits of the continuum, indicating the best point of measurement of the VFQ-25 (Fig. 4). This reveals a clear pattern of better targeting for the more severely affected patients than those with higher functioning.

2. *Ordering of item thresholds*: A total of 12 of the 25 VFQ-25 items had disordered thresholds, which reflected potential problems with the number and/or type of response options in these items (items 4, 12, 13, 15c, 17, and 19 through 25). The top left of Figure 5 shows an item with ordered thresholds (item 3—worry), and the top right of Figure 5 shows an item with disordered thresholds (item 20—stay home). Further details about the remaining items are provided in Appendix Tables A and B in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.10.005.
3. *Item fit validity*: Two of the 25 VFQ-25 items had fit residuals outside of the −2.5 to +2.5 ranges associated with nonsignificant chi-square values, suggesting misfit (item 3—worry and item 4—amount of pain). None of the 25 items was found to have significant chi-square values. In general, the graphical indicators of fit reflected similar fit statistics to those described above, revealing fluctuations in the class intervals in certain ranges of the measurement continuum for some of the items (Fig. 5). The items with the worst deviations were item 3 (worry) and item 4 (amount of pain). These items were found to significantly underdiscriminate across the range of visual functioning.
4. *Item dependency*: Two pairs of items were found to have residual correlations of more than 0.30, suggesting dependency between item scores: item 4 (amount of pain) and item 19 (amount of time without pain), and item 16 (nighttime driving) and item 16a (difficult condition driving).
5. *Reliability*: The estimated Person Separation Index was 0.92, suggesting good reliability.

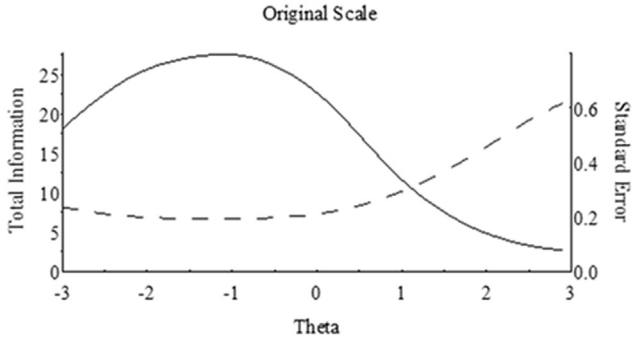

Fig. 3 – IRT: VFQ-25 Test information and standard error. The solid lines correspond to the test information, and the dashed lines correspond to the standard error. IRT, item response theory; VFQ-25, National Eye Institute Visual Functioning Questionnaire.

Overall, the findings from the RMT analysis were mixed in relation to the VFQ-25, and there are two areas requiring further consideration. First, fit and threshold ordering analyses suggest that there is probably more than one clinical concept underpinning the scale and reviewing scale content may help improve validity. Second, targeting analyses suggest that the VFQ-25 may be improved by adding items in the "higher functioning" range of the continuum.

Conclusions

This exercise was designed to compare outcomes when psychometrically evaluating a single PRO instrument in a single data set. Overall, results were similar across the three methods, with the IRT and RMT results providing more diagnostic details to improve the scale.

The CTT approach led to the identification of two problematic items that threaten the validity of the overall scale score, sets of redundant items, and skewed response categories for most items. In addition, the IRT and RMT approaches both identified poor fit for one item, many locally dependent items (threatening reliability and validity), poor targeting (threatening precision of measurement), and disordered thresholds for over half of the response categories (threatening validity). Importantly, both IRT and RMT were able to go beyond the information provided by CTT and begin to shed light on potential causes (e.g., flagging the inappropriate scoring structure for items 20 through 25) and areas for improvement (e.g., response option scoring, item misfit).

Traditional psychometric methods, underpinned by CTT, have been the most commonly used method for over a century and are based on evidence predominantly from correlations and descriptive statistics [5]. Strengths include familiarity, ease of adoption and use, and ability to provide tangible statistics that can be checked against existing criteria. There are, however, four main weaknesses of CTT. First, item-level data are based on ordered counts, not interval-level measurements, despite the fact that interval-level measurement is implied by CTT evaluations. Second, CTT produces findings that are both sample and scale dependent, leading to serious logical drawbacks if the measurement performance of an instrument is affected by the sample it is

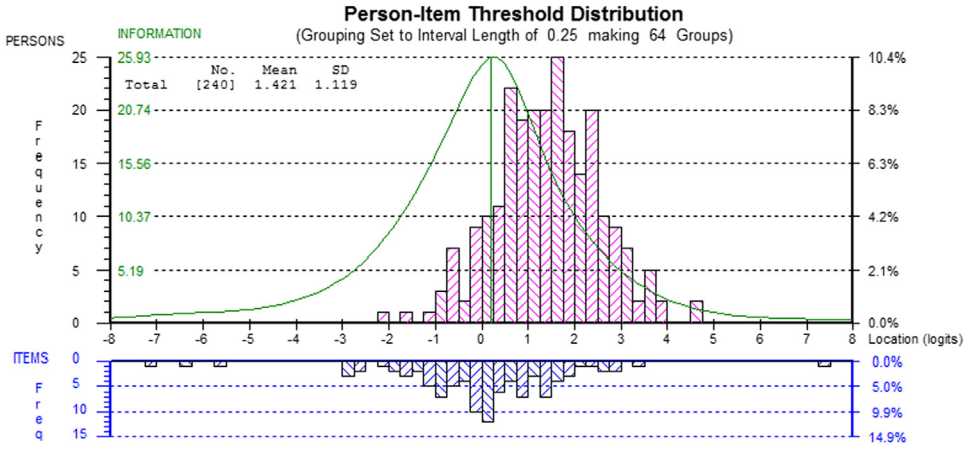

Fig 4. – RMT: VFQ-25—Scale-to-sample targeting (person-item threshold locations spread). The figure shows the person-item thresholds distribution. The x-axis (in logits ranging from –8 to +8) represents the vision functioning construct that the VFQ-25 purports to measure. Better visual functioning increases to the right of the map and decreases to the left (i.e., increasingly positive logits = better functioning; increasingly negative logits = worse functioning). The upper histogram represents the sample distribution of total VFQ-25 person estimates. The lower histogram of striped blue blocks represents the sample distribution of the item thresholds of the 25 items of the same VFQ-25. RMT, Rasch measurement theory; VFQ-25, National Eye Institute Visual Functioning Questionnaire.

supposed to be measuring and vice versa. Third, missing data cannot be handled easily, and conservative rules to set total scores to missing when a certain number of items are missing unnecessarily ignore the information contained in the nonmissing items [45]. Finally, the standard error of measurement around individual patients' scores is assumed to be a constant value regardless of the person's location on the range of a scale, but it is counterintuitive that patients' scores at the extremes of the scale (floor and ceiling) have the same level of precision as those scores in the center of the scale (where most of the items would lie).

Modern psychometric approaches (i.e., IRT and RMT) represent a logical progression from CTT because they use more

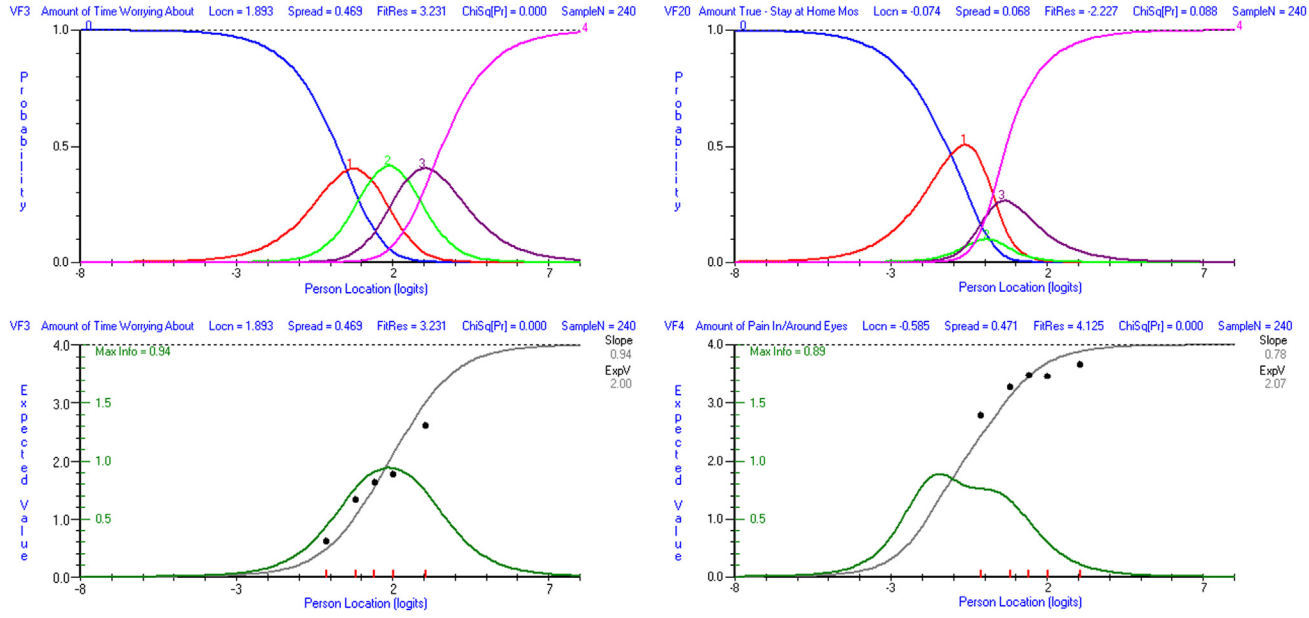

Fig. 5 – RMT: VFQ-25—RMT category probability curves and item fit. The first row of plots in the figure shows category probability curves for two example VFQ-25 items, and the second row of plots shows item characteristic curves. The x-axis (in logits ranging from –8 to +8) represents the vision functioning construct that the VFQ-25 purports to measure. Better visual functioning increases to the right of the map and decreases to the left (i.e., increasingly positive logits = better functioning; increasingly negative logits = worse functioning). In the first row of plots, the different colored lines represent the response categories for each of the items, and the y-axis shows the probability of endorsing the response categories that are specific to each item. In the second row of plots, the y-axis shows the expected value as predicted by the Rasch model. The line represents the expected scores as predicted by the Rasch model. The black dots represent class intervals (mean person estimates). The closer the dots are to the line, the better the fit to the Rasch model. FitRes, fit residual; Locn, location; Pr, probability; RMT, Rasch measurement theory; VFQ-25, National Eye Institute Visual Functioning Questionnaire.

sophisticated models and techniques. Although IRT and RMT are both item response models with a commonality in the structure of the mathematical models that are used, they approach social measurement from different starting points. The main difference is in how the two methods approach measurement and the evaluation of an item set. IRT prioritizes the data and aims to find the item response model that best explains the data. RMT prioritizes the Rasch model, and if data do not fit, the hypothesis requires revisiting. The reasons underlying these stances are expounded in more detail elsewhere (e.g., Andrich [17]). We would suggest that it is essential to understand these differences for researchers deciding which approach to adopt, as well as to know that they each require a complex advanced level of mathematical understanding and unique software.

It is important to note that none of the approaches provide truly sample-free estimates. If the Rasch model is used (called the 1PL in IRT), sample and scale distribution-free estimates can be obtained [6]. In the IRT paradigm, as additional parameters are added (e.g., 2PL and 3 PL) through the subsequent IRT models, however, the parameter estimates are sample dependent within the same model up to linear transformation [29].

The choice of the psychometric approach depends on a number of factors. The researchers should select methods they are comfortable performing because the current availability of user-friendly "black box" statistical software makes many psychometric methods more accessible than they have been in the past, but blind application of these methods can result in erroneous conclusions. The intended audience must also be considered. If the instrument is being developed for descriptive purposes and on a restricted budget, a cursory examination of the CTT-based psychometric properties may be all that is possible. We would propose that any level of psychometric analysis is better than none for these sorts of cases. In a high-stakes situation, however, such as the development of a PRO instrument for consideration in pharmaceutical labeling, a thorough psychometric evaluation must be performed using methods considered appropriate by the regulatory body, with both qualitative and quantitative results factoring in decisions. We suggest that researchers use IRT or RMT where appropriate for an evaluation. Simple problems with respect to missing data and floor/ceiling effects, however, are easily identified by CTT, so do not overlook the value of CTT. For example, CTT results identified more problems with the "driving at night" item than did the IRT results; the item had 50% missing responses and was flagged by CTT, but based on the IRT results, this item was selected as the candidate item to keep from the set of driving items because it was viewed as the most informative.

The main limitation of this study is that the consistency of results across the three psychometric methods may have been influenced by the instrument and data set chosen. The VFQ-25 is a well-known legacy measure that, like many of its time, was developed using what can now be considered a limited development process but 10 years ago was gold standard. Thus, it is not surprising that problems exist and were consistently identified. It would therefore be valuable to replicate our efforts here in other instruments to see whether any of the methods identify issues that other methods do not. An additional limitation of this study was the focus on the three methods in isolation versus as a component of a more extensive psychometric evaluation. Furthermore, we have not exhausted all aspects of these approaches. A typical evaluation would include an evaluation of the dimensionality of an item set based on methods such as exploratory factor analysis or confirmatory factor analysis. The authors do not wish to imply that other aspects of item and scale evaluation should be omitted. Our attempt, however, was to apply typical analyses specific to each of the three paradigms to a common data set.

Although direct comparisons of psychometric methods in the clinical literature are very rare, researchers from different paradigms working together in a single project are even rarer. Ultimately, the aim of all psychometricians working in outcomes measurement settings, regardless of approach, is to improve the available methodology, including the combination of qualitative and quantitative information in high-stakes evaluations. We hope that our unique collaborative effort will encourage others to do the same.

Source of financial support: Novartis supported this study.

Supplemental Material

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2014.10.005 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Patient-Centered Outcomes Research Institute. Patient-centered outcomes research. 2012. Available from: http://www.pcori.org/patient-centered-outcomes-research/. [Accessed March 23, 2012].
[2] Wilson I, Cleary P. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. JAMA 1995;273:59–65.
[3] Gnanasakthy A, Mordin M, Clark M, et al. A review of patient-reported outcome labels in the United States: 2006 to 2010. Value Health 2012;15:437–42.
[4] US Department of Health and Human Services. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. December 2009. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. [Accessed June 13, 2014].
[5] Novick MR. The axioms and principal results of classical test theory. J Math Psychol 1966;3:1–18.
[6] Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Welsley Publishing Company, 1968.
[7] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1980.
[8] Spearman C. "General intelligence," objectively determined and measured. Am J Psychol 1904;15:201–92.
[9] Lord FM. Applications of Item Response Theory to Practical Testing Problems. Mahwah, NJ: Erlbaum, 1980.
[10] Thissen D, Wainer H. Test Scoring. Mahwah, NJ: Erlbaum, 2001.
[11] Nunnally JC, Bernstein IH. Psychometric Theory. (3rd ed.. New York NY: McGraw-Hill, 1994.
[12] Traub R. Classical test theory in historical perspective. Educ Meas Issues Pract 1997;16:8–14.
[13] Cappelleri JC, Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36:648–62.
[14] Wright B. A history of social science and measurement. Educ Meas Issues Pract 1997;16:33–52.
[15] Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;42:I7–16.
[16] Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. Health Technology Assess 2009;13: iii, ix–x, 1–177.
[17] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoeconomics Outcomes Res 2011;11:571–85.
[18] Cano S, Hobart J. The problem with health measurement. Patient Prefer Adherence 2011;5:279–90.
[19] Andrich D. The legacies of R. A. Fisher and K. Pearson in the application of the polytomous Rasch model for assessing the empirical ordering of categories. Educ Psychol Meas 2013;73:553–80.
[20] Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR, eds., Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.
[21] Bock RD. A brief history of item response theory. Educ Meas Issues Pract 1997;16:21–33.
[22] Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005;28:264–82.

[23] Thissen D, Steinberg L. Item response theory. In: Millsap R, Maydeu-Olivares A, eds., The Sage Handbook of Quantitative Methods in Psychology. London: Sage, 2009.
[24] Nguyen TH, Han H-R, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7:23–35.
[25] The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 1998;46:1569–85.
[26] Aaronson N, Alonso J, Burnam A, et al. Assessing health status and quality of life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
[27] Mokkink L, Terwee C, Patrick D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539–49.
[28] Wright B, Masters G. Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press, 1982.
[29] Massof RW. Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. Ophthalmic Epidemiol 2011;18:1–19.
[30] Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:294–334.
[31] Mangione CM, Lee PP, Gutierrez PR, et al. Development of the 25-item National Eye Institute Visual Function Questionnaire. Arch Ophthalmol 2001;119:1050–8.
[32] Mitchell P, Bandello F, Schmidt-Erfurth U, et al. The RESTORE study: ranibizumab monotherapy or combined with laser versus laser monotherapy for diabetic macular edema. Ophthalmology 2011;118:615–25.
[33] Mitchell P, Bressler N, Tolley K, et al. Patient-reported visual function outcomes improve after ranibizumab treatment in patients with vision impairment due to diabetic macular edema: randomized clinical trial. JAMA Ophthalmol 2013;131:1339–47.
[34] Orr P, Rentz AM, Margolis MK, et al. Validation of the National Eye Institute Visual Function Questionnaire-25 (NEI VFQ-25) in age-related macular degeneration. Invest Ophthalmol Vis Sci 2011;52:3354–9.
[35] Khadka J, McAlinden C, Pesudovs K. Validation of the National Eye Institute Visual Function Questionnaire-25 (NEI VFQ-25) in age-related macular degeneration. Invest Ophthalmol Vis Sci 2012;53:1276.
[36] Marella M, Pesudovs K, Keeffe JE, et al. The psychometric validity of the NEI VFQ-25 for use in a low-vision population. Invest Ophthalmol Vis Sci 2010;51:2878–84.
[37] Bulmer MG. Principles of Statistics. Mineola, NY: Dover Publications, 1979.
[38] Cai L, Du Toit SHC, Thissen D. IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling [Computer software]. Chicago: Scientific Software International, 2012.
[39] Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychom Monogr 1969;17:19–36.
[40] Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Measure 2000;24:50–64.
[41] Orlando M, Thissen D. Further examination of the performance of S-X^2, an item fit index for dichotomous item response theory models. Appl Psychol Measure 2003;27:289–98.
[42] Chen WH, Thissen D. Local dependence indices for item pairs using item response theory. J Educ Behav Stat 1997;22:265–89.
[43] Nunnally JC. Psychometric Theory. (2nd ed). New York: McGraw-Hill, 1978.
[44] Andrich D. An index of person separation in latent trait theory, the traditional KR20 index, and the Guttman scale response pattern. Educ Psychol Res 1982;9:95–104.
[45] Coon CD, Williams VSL, Nelson LM, Price MA. Determining missing data rules for patient-reported outcomes: alpha-if-item-deleted. Poster presented at the 15th Annual Meeting of the International Society for Pharmacoeconomics and Outcomes Research, May 17, 2010, Atlanta, GA.