UCLA
UCLA Previously Published Works

Title

Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures

Permalink

https://escholarship.org/uc/item/5zw8z6wk

Journal

Clinical Therapeutics, 36(5)

ISSN

0149-2918

Authors

Cappelleri, Joseph C
Lundy, J Jason
Hays, Ron D

Publication Date

2014-05-01

DOI

10.1016/j.clinthera.2014.04.006

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Commentary

Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures

Joseph C. Cappelleri, PhD[1]; J. Jason Lundy, PhD[2]; and Ron D. Hays, PhD[3]

[1]Pfizer Inc, Groton, Connecticut; [2]Critical Path Institute, Tucson, Arizona; and [3]Division of General Internal Medicine & Health Services Research, University of California at Los Angeles, Los Angeles, California

ABSTRACT

Background: The US Food and Drug Administration's guidance for industry document on patient-reported outcomes (PRO) defines *content validity* as "the extent to which the instrument measures the concept of interest" (FDA, 2009, p. 12). According to Strauss and Smith (2009), construct validity "is now generally viewed as a unifying form of validity for psychological measurements, subsuming both content and criterion validity" (p. 7). Hence, both qualitative and quantitative information are essential in evaluating the validity of measures.

Methods: We review classical test theory and item response theory (IRT) approaches to evaluating PRO measures, including frequency of responses to each category of the items in a multi-item scale, the distribution of scale scores, floor and ceiling effects, the relationship between item response options and the total score, and the extent to which hypothesized "difficulty" (severity) order of items is represented by observed responses.

Results: If a researcher has few qualitative data and wants to get preliminary information about the content validity of the instrument, then descriptive assessments using classical test theory should be the first step. As the sample size grows during subsequent stages of instrument development, confidence in the numerical estimates from Rasch and other IRT models (as well as those of classical test theory) would also grow.

Conclusion: Classical test theory and IRT can be useful in providing a quantitative assessment of items and scales during the content-validity phase of PRO-measure development. Depending on the particular type of measure and the specific circumstances, the classical test theory and/or the IRT should be considered to help maximize the content validity of PRO measures. (*Clin Ther*. 2014;36:648–662) © 2014 Elsevier HS Journals, Inc. All rights reserved.

Key words: classical test theory, content validity, item response theory, patient-reported outcomes, scale development.

INTRODUCTION

The publication of the US Food and Drug Administration's guidance for industry on patient-reported outcomes (PRO)[1] has generated discussion and debate on the methods used for developing, and establishing the content validity of, PRO instruments. The guidance outlines the information that the FDA considers when evaluating a PRO measure as a primary or secondary end point to support a claim in medical product labeling. The PRO guidance highlights the importance of establishing evidence of *content validity*, defined as "the extent to which the instrument measures the concept of interest" (p. 12).[1]

Accepted for publication April 9, 2014.
http://dx.doi.org/10.1016/j.clinthera.2014.04.006
0149-2918/$ - see front matter

© 2014 Elsevier HS Journals, Inc. All rights reserved.

Scan the QR Code with your phone to obtain FREE ACCESS to the articles featured in the Clinical Therapeutics topical updates or text GS2C65 to 64842. To scan QR Codes your phone must have a QR Code reader installed.

Content validity is the extent to which an instrument covers the important concepts of the unobservable, or latent, attribute (eg, depression, anxiety, physical functioning, self-esteem) that the instrument purports to measure. It is the degree to which the content of a measurement instrument is an adequate reflection of the construct being measured. Hence, qualitative work with patients is essential to ensure that a PRO instrument captures all of the important aspects of the concept from the patient's perspective.

Two reports from the International Society of Pharmacoeconomics and Outcomes Research Good Research Practices Task Force[2,3] detail the qualitative methodology and 5 steps that should be employed to establish content validity of a PRO measure: (1) determine the context of use (eg, medical product labeling); (2) develop the research protocol for qualitative concept elicitation and analysis; (3) conduct the concept elicitation interviews and focus groups; 4) analyze the qualitative data; and (5) document concept development, elicitation methodology, and results. Essentially, the inclusion of the entire range of relevant issues in the target population embodies adequate content validity of a PRO instrument.

Although qualitative data from interviews and focus groups with the targeted patient sample are necessary to develop PRO measures, qualitative data alone are not sufficient to document the content validity of the measure. Along with qualitative methods, quantitative methods are needed to develop PRO measures with good measurement properties. Quantitative data gathered during earlier stages of instrument development can serve as: (1) a barometer to see how well items address the entire continuum of the targeted concept of interest; (2) a gauge of whether to go forward with psychometric testing; and (3) a meter to mitigate risk related to Phase III signal detection and interpretation.

Specifically, quantitative methods can support the development of PRO measures by addressing several core questions of content validity: What is the range of item responses relative to the sample (distribution of item responses and their endorsement)?; Are the response options used by patients as intended?; Does a higher response option imply a greater health problem than does a lower response option?; and What is the distance between response categories in terms of the underlying concept?

Also relevant is the extent to which the instrument reliably assesses the full range of the target population (scale-to-sample targeting), ceiling or floor effects, and the distribution of the total scores. Does the item order with respect to disease severity reflect the hypothesized item order? To what extent do item characteristics relate to how patients rank the items in terms of their importance or bother?

This article reviews the classical test theory and the item response theory (IRT) approaches to developing PRO measures and to addressing these questions. These content-based questions and the 2 quantitative approaches to addressing them are consistent with construct validity, now generally viewed as a unifying form of validity for psychological measurements, subsuming both content and criterion validity.[4] The use of quantitative methods early in instrument development is aimed at providing descriptive profiles and exploratory information about the content represented in a draft PRO instrument. Confirmatory psychometric evaluations, occurring at the later stages of instrument development, should be used to provide more definitive information regarding the measurement characteristics of the instrument.

CLASSICAL TEST THEORY

Classical test theory is a conventional quantitative approach to testing the reliability and validity of a scale based on its items. In the context of PRO measures, classical test theory assumes that each observed score (X) on a PRO instrument is a combination of an underlying true score (T) on the concept of interest and nonsystematic (ie, random) error (E). Classical test theory, also known as *true-score theory,* assumes that each person has a true score, T, that would be obtained if there were no errors in measurement. A person's *true score* is defined as the expected score over an infinite number of independent administrations of the scale. Scale users never observe a person's true score, only an observed score, X. It is assumed that observed score (X) = true score (T) + some error (E).

True scores quantify values on an *attribute of interest,* defined here as the underlying concept, construct, trait, or ability of interest (the "thing" intended to be measured). As values of the true score increase, responses to items representing the same concept should also increase (ie, there should be a monotonically increasing relationship between true scores and item scores), assuming that item responses are coded so that higher responses reflect more of the concept.

It is also assumed that random errors (ie, the difference between a true score and a set of observed scores in the same individual) found in observed scores are normally distributed and, therefore, that the expected value of such random fluctuations (ie, mean of the distribution of errors over a hypothetical infinite number of administrations in the same subject) is taken to be zero. In addition, random errors are assumed to be uncorrelated with a true score, with no systematic relationship between a person's true score and whether that person has positive or negative errors.

Descriptive Assessment

In the development of a PRO measure, the means and SDs of the items can provide fundamental clues about which items are useful for assessing the concept of interest. Generally, the higher the variability of the item scores and the closer the mean score of the item is to the center of its distribution (ie, median), the better the item will perform in the target population.

In the case of *dichotomous* items, scored "0" for one response and "1" for the other (eg, *no* = 0 *and yes* = 1), the proportion of respondents in a sample who select the response choice scored "1" is equivalent to the mean item response. If a particular response option (typically given a response choice scored "1") represents an affirmative response or the presence of what is asked, the item is said to be *endorsed*. Items for which everyone gives the same response are uninformative because they do not differentiate between individuals. In contrast, dichotomous items that yield about an equal number of people (50%) selecting each of the 2 response options provide the best differentiation between individuals in the sample overall.

For items with *ordinal* response categories, which have >2 categories, an equal or a uniform distribution across response categories yields the best differentiation. Although ideal, such a uniform distribution is typically difficult to obtain (unless the researcher makes it a direct part of the sampling frame during the design stage) as it depends in part on the distribution of the sampled patients, which is outside the full control of the researcher.

Item difficulty, which is taken from educational psychology, may or may not be an apt term in health care settings. In this article, we equate *item difficulty* with *item severity,* and we use the 2 terms interchangeably. The more suitable and interpretable term depends on the particular health care application. Item difficulty, or severity, can be expressed on a z-score metric by transforming the proportion endorsed using the following formula: $z = ln[p/(1 - p)]/1.7$, where z scores come from a standardized normal distribution (with a mean of 0, SD of 1), ln represents the natural logarithm, p represents the probability endorsement, and 1.7 is a scaling factor for a normal distribution. The z scores of the items can be ordered so that, for instance, the items with higher z scores are considered more difficult relative to the other items. For items with ordinal response categories, adjacent categories can be combined meaningfully to form a dichotomous indicator for the purpose of examining item severity (difficulty).

Item Discrimination

The more an item discriminates among individuals with different amounts of the underlying concept of interest, the higher the *item-discrimination index*.[5] The discrimination index can be applied with a dichotomous item response or an ordinal item response made dichotomous (by combining adjacent categories meaningfully to form a dichotomous indicator).

The *extreme group method* can be used to calculate the discrimination index using the following 3 steps. Step 1 is to partition respondents who have the highest and lowest overall scores on the overall scale, aggregated across all items, into upper and lower groups. The upper group can be composed of the top x% (eg, 25%) of scores on the scale, and the lower group can be composed of the bottom x% (eg, 25%) of scores on the scale. Step 2 is to examine each item and determine the proportion of individual respondents in the sample who endorse or respond to each item (or to a particular category or adjacent category groups of an

Table I. Illustration example of using the item-discrimination index.

Item	Proportion Endorsed for Upper Group	Proportion Endorsed for Lower Group	Item-Discrimination Index
1	0.90	0.10	0.80
2	0.85	0.20	0.65
3	0.70	0.65	0.05
4	0.10	0.70	–0.60

item) in the upper and lower groups. Step 3 is to subtract the pair of proportions noted in Step 2. The higher this item-discrimination index, the more the item discriminates. For example, if 60% of the upper group and 25% of the lower group endorse a particular item in the scale, the item-discrimination index for that item would be calculated as: 0.60 – 0.25 = 0.35. It is useful to compare the discrimination indexes of each of the items in the scale, as illustrated in Table I. In this example, item 1 provides the best discrimination, item 2 provides the next best discrimination, and items 3 and 4 are poor discriminators.

Another indicator of item discrimination is how well an item correlates with the sum of the remaining items on the same scale or domain, or the *corrected item-to-scale correlation* ("corrected" because the sum or total score does not include that item). It is best to have relatively "large" corrected item-to-scale correlations (eg, ≥0.37 according to Cohen's rule of thumb).[6] A low corrected item-to-scale correlation indicates that the item is not as closely associated with the scale relative to the rest of the items in the scale.

The response categories of an item can be assessed by analyzing the *item response curve,* which is produced descriptively in classical test theory by plotting the percentage of subjects who choose each response option on the *y*-axis and the total score, expressed as such or as a percentile or other metric, on the *x*-axis. Figure 1 provides an illustration. Item 1 is an equally good discriminator across the continuum of the attribute (the concept of interest). Item 2 discriminates better at the lower end than at the upper end of the attribute. Item 3 discriminates better at the upper end, especially between 70th and 80th percentiles.

A *difficulty-by-discrimination graph,* shown in Figure 2, depicts how well the items in a scale span across the range of difficulty (or severity), along with the how well each item represents the concept. The data shown are 16 items included in a 50-item test administered to 5 students. (Thirty-three of the items in the test were answered correctly by all 5 students, and 1 item was answered incorrectly by all 5 students.) Figure 2 shows the 16 items by sequence number in the test. For example, the third item in the test, shown in the upper left of Figure 2, is labeled "3." This item ("Which of the following could be a patient-reported measure?") was answered correctly by 4 of the 5 students (80%) and had a corrected item-to-scale correlation of –0.63 ("corrected" to remove that item from the total scale score). This item had a negative correlation with the total scale score because the student who had the best test score was the only student of the 5 who answered the question incorrectly. Items 10 and 27 both were answered correctly by 2 of the 5 students (40%), and each had an item-to-scale correlation of 0.55. The "easiest" or "least severe" items (100% of students got them right) and the "hardest" or "most

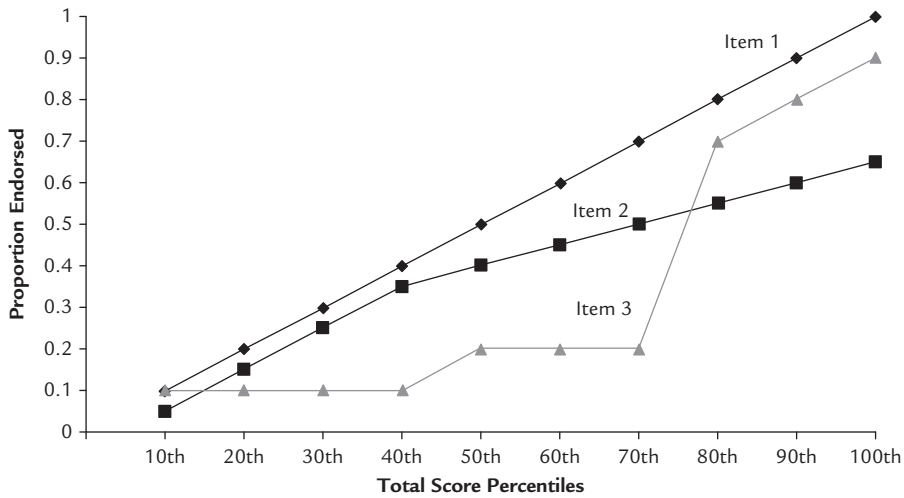

Figure 1. Illustrative item response curves.

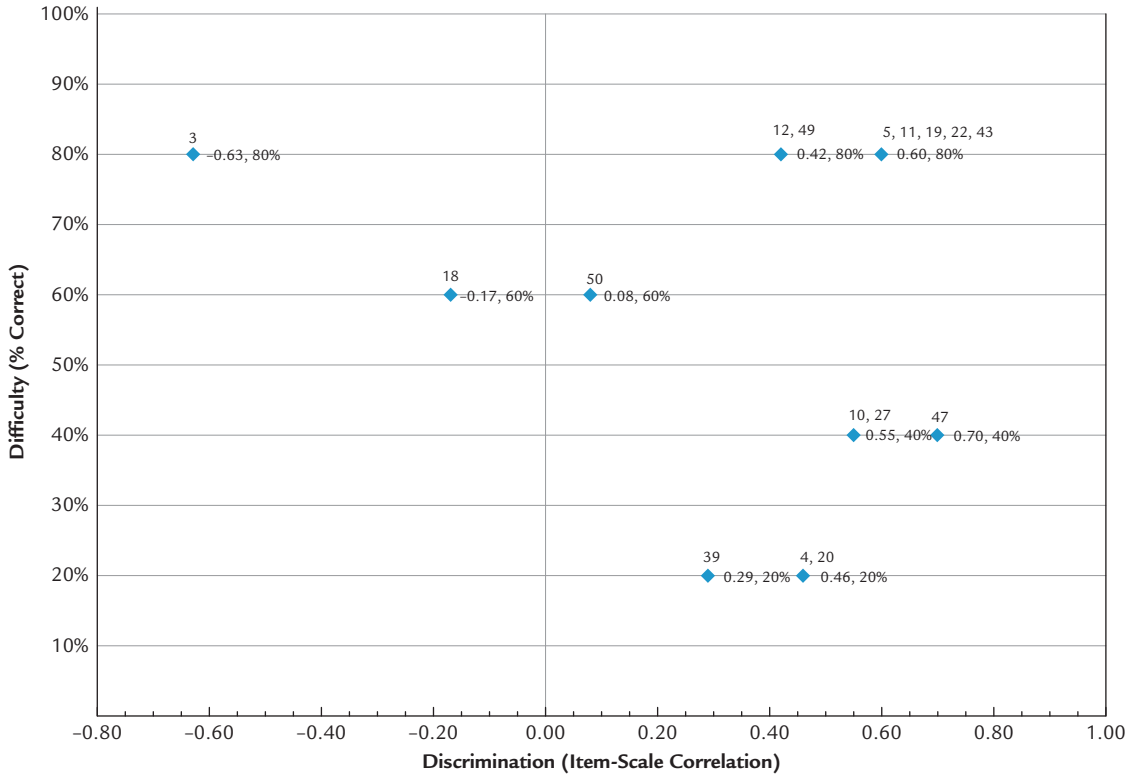

Figure 2. Difficulty-by-discrimination graph. Numbers in the figure represent sequence in which the item was administered in a 50-item test given to 5 students. Sixteen items are shown in the figure. Thirty-three items that all of the students answered correctly and 1 item that no students answered correctly are not shown.

severe" item (0% of students got it right) are not shown in Figure 2. The range-of-difficulty estimates for the other items are limited by the small sample of students, but Figure 2 shows that 8 of the items were easier items (80% correct), 2 items were a little harder (60% correct), 3 items were even harder (40% correct), and 3 other items were among the hardest (20% correct).

Dimensionality

To evaluate the extent to which the items measure a hypothesized concept distinctly, item-to-scale correlations on a particular scale (corrected for item overlap with the total scale score) can be compared to correlations of those same items with other scales (either subscales within the same PRO measure or different scales from different PRO measures). This approach has been referred to as *multitrait scaling analysis* and can be implemented, for example, using a SAS macro (SAS Institute Inc, Cary, North Carolina).[7] Although some users of the methodology suggest that it evaluates item convergent and discriminant validity, we prefer "item convergence within scales" and "item discrimination across scale," because one learns that items sort into different scales ("bins") but that the validity of the scales per se is still unknown.

Factor analysis is a statistical procedure analogous to multitrait scaling.[8] In *exploratory factor analysis,* there is uncertainty as to the number of factors being measured; the results of the analysis are used to help identify the number of factors. Exploratory factor analysis is suitable for generating hypotheses about the structure of the data. In addition, it can help in further refining a PRO instrument by revealing what items may be dropped from the instrument because they contribute little to the presumed underlying factors. Whereas exploratory factor analysis explores the patterns in the correlations of items

(or variables), *confirmatory factor analysis* (which is appropriate for later stages of PRO development) tests whether the variance–covariance of items conforms to an anticipated or expected scale structure given in a particular research hypothesis. Although factor analysis is not necessarily connected to content validity per se, it can be useful for testing the conceptual framework of the items mapping to the hypothesized underlying factors.

Reliability

Reliability is important in the development of PRO measures, including for content validity. Validity is limited by reliability. If responses are inconsistent (unreliable), it necessarily implies invalidity as well (note that the converse may not be true: consistent responses do not necessarily imply valid responses).[9] Although we are primarily concerned with validity in early scale development, reliability is a necessary property of the scores produced by a PRO instrument and it is important to consider in early scale development. *Reliability* refers to the proportion of variance in a measure that can be ascribed to a common characteristic shared by the individual items, whereas *validity* refers to whether that characteristic is actually the one intended.

Test–retest reliability, which can apply to both single-item and multi-item scales, reflects the reproducibility of scale scores on repeated administrations over a period during which the respondent's condition did not change. As a way to compute test–retest reliability, the *kappa statistic* can be used for categorical responses, and the *intraclass correlation coefficient* can be used for continuous responses (or responses taken as such).

Further, having multiple items in a scale increases its reliability. In multi-item scales, a common indicator of scale reliability is *Cronbach coefficient alpha,* which is driven by 2 elements: (1) the correlations between the items and (2) the number of items in the scale. In general, the reliability of a measure equals the proportion of total variance among its items that is due to the latent variable and is thus considered *communal* or *shared variance*.

The greater the proportion of shared variation, the more the items share in common and the more consistent they are in reflecting a common true score. The *covariance-based formula* for coefficient alpha expresses such reliability while adjusting for the number of items contributing to the prior calculations on the variances. The *corresponding correlation–based formula,* an alternative expression, represents coefficient alpha as the mean inter-item correlation among all pairs of items after adjustment for the number of items.

Sample-Size Considerations

In general, different study characteristics affect sample-size considerations, such as the research objective, type of statistical test, sampling heterogeneity, statistical power or level of confidence, error rates, and the type of instrument being tested (eg, the number of items and the number of categories per item). In the quantitative component of content validity, a stage considered exploratory, a reliable set of precise values for all measurement characteristics is not expected and, as such, formal statistical inferences are not recommended. Consequently, sample-size adequacy in this early stage, which should emphasize absolute and relative directionality, does not have the same level of importance as it would in later (and especially confirmatory) phases of PRO-instrument development, regardless of the methodology employed.

Nonetheless, sample sizes based on classical test theory should be large enough for the descriptive and exploratory pursuit of meaningful estimates from the data. Although it is not appropriate to give 1 number for sample size in all such cases, starting with a sample of 30 to 50 subjects may be reasonable in many circumstances. If no clear trends emerge, more subjects may be needed to observe any noticeable patterns. An appropriate sample size depends on the situation at hand, such as the number of response categories. An 11-point numerical rating scale, for instance, may not have enough observations in the extreme categories and thus may require a larger sample size. In addition to increasing the sample size, another way to have a more even level of observations across categories of a scale is to plan for it at the design stage by recruiting individuals who provide sufficient representation across the response categories.

Sample sizes for more rigorous quantitative analyses, at the later stages of psychometric testing, should be large enough to meet a desired level of measurement precision or SE.[10] With sample sizes of 100, 200, 300, and 400, the SEs around a correlation are ~0.10, 0.07, 0.06, and 0.05, respectively. Various recommendations have been given for exploratory factor analyses. One recommendation is to have at least 5 cases per item and a minimum of 300 cases.[11] Another rule of thumb is to enlist a sample size of at

least 10-fold the number of items being analyzed, so a 20-item questionnaire would require at least 200 subjects. However, adequate sample size is directly dependent on the properties of the scale itself, rather than on rules of thumb. A poorly defined factor (eg, one with not enough items) or weakly related items (low factor loading) may require substantially more individuals for obtaining precise estimates.[12]

For confirmatory factor analysis, rules of thumb have been offered about the minimum number of subjects per each parameter to be estimated (eg, at least 10 subjects per parameter).[13] The same caution given about such rules of thumb for exploratory factor analysis also applies to confirmatory factor analysis; that is, if a measure is to be used in a specific subgroup (eg, Asian-American subjects), then a sufficient sample size is needed to represent that subgroup. Statistical power and sample sizes for confirmatory factor analysis are explained in more detail elsewhere.[13]

In some situations (eg, if large patient accrual is not feasible or if responses are diverse or heterogeneous in spanning across item categories), a smaller sample size might be considered sufficient. In these situations, analytical methods can include simple descriptive statistics (item-level means, SEs and counts, and correlations between items) for the items and subscales of a PRO measure. Replication of psychometric estimates is needed either by a sufficiently large and representative sample that can be split into 2 subsamples for cross-validation or 2 samples of sufficient sample size. One sample is used to explore the properties of the scale, and the second sample is used to confirm the findings of the first sample. If the results of the 2 samples are inconsistent, then psychometric estimates from another sample may be required to establish the properties of the measure.

ITEM RESPONSE THEORY

Item response theory (IRT) is a collection of measurement models that attempt to explain the connection between observed item responses on a scale and an underlying construct. Specifically, IRT models are mathematical equations describing the association between subjects' levels on a latent variable and the probability of a particular response to an item, using a nonlinear monotonic function.[14] As in classical test theory, IRT requires that each item be distinct from the others, yet similar and consistent with them in reflecting all important respects of the underlying attribute or construct. Item parameters in IRT are estimated directly using logistic models instead of proportions (difficulty or threshold) and item-to-scale correlations (discrimination). There are a number of IRT models, which vary in number of parameters (1-, 2-, or 3-parameter models) and in whether they handle dichotomous-only items or polytomous items more generally (Table II).

In the simplest case, item responses are evaluated in terms of a single parameter, difficulty (severity). For a dichotomous item, the difficulty, or severity, parameter indicates the level of the attribute (eg, the level of physical functioning) at which a respondent has a 50% likelihood of endorsing the dichotomous item. In the present article, without loss of generality, items with higher levels of difficulty or severity are those that require higher levels of health—for instance, running as opposed to walking for physical functioning. Items and response options can be written in the other direction so that "more difficult" requires "worse health," but here the opposite is assumed.

For a polytomous item, the meaning of the difficulty or severity parameter depends on the model used and represents a set of values for each item. In a *graded-response model,* a type of 2-parameter model that allows for item discrimination as well as difficulty to vary across items, the difficulty parameter associated with a particular category k of an item reflects the level of the attribute at which patients have 50% likelihood of scoring a category lower than k versus category k or higher. In a *partial-credit model,* a generalization of the 1-parameter (Rasch) IRT dichotomous model, in which all items have equal discrimination, the difficulty parameter is referred to as the *threshold parameter* and reflects the level of the attribute at which the probability of a response in either 1 of 2 adjacent categories is the same.

A 1-parameter (Rasch) IRT model for dichotomous items can be written as follows:

$$P_i(X=1|\Theta)=[e^{(\Theta-b_i)}]/[1+e^{(\Theta-b_i)}]$$

where $P_i(X = 1|\Theta)$ is the probability that a randomly selected respondent on the latent trait with level Θ (the Greek letter theta) will endorse item i, and b_i is the item difficulty (severity) parameter. In the 1-parameter IRT model, each item is assumed to have the same amount of item discrimination.

In a 2-parameter IRT model, an item-discrimination parameter is added to the model. A 2-

Table II. Common item response theory models applied to patient-reported outcomes.

Model	Item Response Format	Model Characteristics
1-Parameter (Rasch) logistic	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
2-Parameter logistic	Dichotomous	Discrimination and threshold parameters vary across items.
Graded response	Polytomous	Ordered responses. Discrimination varies across items.
Nominal	Polytomous	No prespecified item order. Discrimination varies across items.
Partial credit (Rasch model)	Polytomous	Discrimination and power constrained to be equal across items.
Rating scale (Rasch model)	Polytomous	Discrimination equal across items. Item-threshold steps equal across items.
Generalized partial credit	Polytomous	Variation of partial-credit model with discrimination varying across items.

parameter model for a dichotomous item can be written as follows:

$$P_i(X=1|\Theta)=[e^{Da_i(\Theta-b_i)}]/[1+e^{Da_i(\Theta-b_i)}]$$

where D is a scaling constant ($D = 1.7$ represents the normal ogive model), a_i is the discrimination parameter, and the other variables are the same as those in the 1-parameter model. An important feature of the 2-parameter model is that the distance between an individual's trait level and an item's severity has a greater impact on the probability of endorsing highly discriminating items than on less discriminating items. In particular, more discriminating items provide more information (than do less discriminating items) and even more so if a respondent's level on the latent attribute is closer to an item's location of severity.

Item-Characteristic Curve

The *item-characteristic curve* (ICC) is the fundamental unit in IRT and can be understood as the probability of endorsing an item (for a dichotomous response) or responding to a particular category of an item (for a polytomous response) for individuals with a given level of the attribute. In the latter case, the ICC is sometimes referred to as a *category-response curve*. Depending on the IRT model used, these curves indicate which items (or questions) are more difficult and which items are better discriminators of the attribute.

For example, if the attribute were mental health, the person with better mental health (here assumed to have higher levels of Θ) would be more likely to respond favorably to an item that assesses better mental health (an item with a higher level of "difficulty" needed to achieve that better state of mental health). If an item were a good discriminator of mental health, the probability of a positive response to this item (representing better mental health) would increase more rapidly as the level of mental health increases (larger slope of the ICC); given higher levels of mental health, the (conditional) probability of a positive response would increase noticeably across these higher levels. The various IRT models, which are variations of logistic (ie, nonlinear) models, are simply different mathematical functions for describing ICCs as the relationship of a person's level on the attribute and an item's characteristics (eg, difficulty, discrimination) with the probability of a specific response on that item measuring the same attribute.

Category-Response Curves

In IRT models, a function, which is analogous to an ICC for a dichotomous response, can be plotted for each category of an item with >2 response categories (ie, polytomous response scale). Such category-response curves help in the evaluation of response options for each item by displaying the relative position of each category along the underlying continuum of the concept being measured. The ideal category-response curve is characterized by each response category being most likely to be selected for some seg-

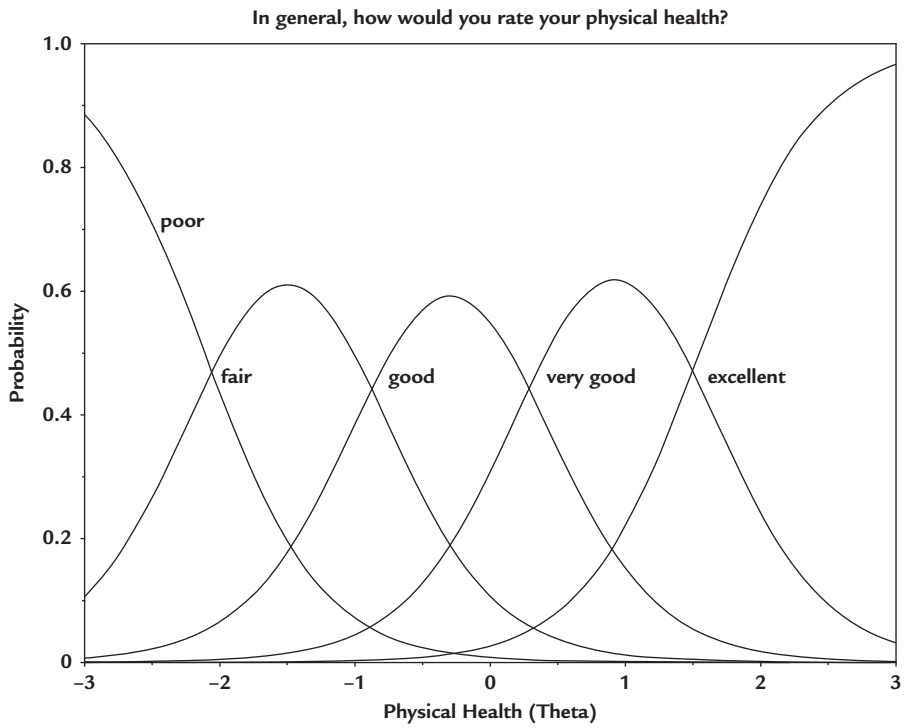

Figure 3. Category-response curves for PROMIS (Patient-Reported Outcomes Measurement Information System) Physical Health summary score from grade response model. The labels indicate the five response options (poor, fair, good, very good, excellent) for the item. The probability of picking each of the response option by estimated physical health (theta [Θ]) was based on data from the PROMIS project.[15]

ment of the underlying continuum of the attribute (the person's location on the attribute), with different segments corresponding to the hypothesized rank order of the response options in terms of the attribute or concept.

Figure 3 shows an example category-response curve for an item in the PROMIS (Patient-Reported Outcomes Measurement Information Measurement System, a National Institutes of Health initiative to utilize IRT to develop item banks assessing patient-reported outcomes) 4-item General Health scale: "In general, how would you rate your physical health?"[15] The item has 5 response options: poor, fair, good, very good, and excellent. The x-axis of Figure 3 shows the estimated Physical Health score (Θ) depicted on a z-score metric, with a more positive score representing better physical health. The y-axis shows the probability of selecting each response option. Figure 3 shows that the response options for the item are monotonically related to physical health as expected, and that each response option is most likely to be selected at some range of the underlying construct (ie, physical health).

Item Information

Item information provides an assessment of the precision of measurement of an item in distinguishing among subjects across different levels of the underlying concept or attribute being measured (Θ); higher information implies more precision. Item information depends on the item parameters. In the 2-parameter dichotomous logistic model, the item-information function [$I(\Theta)$] for item i at a specific value of Θ is equal to $(a_i^2)\ P_i(1 - P_i)$, where P_i is the proportion of people with a specific amount of the attribute who endorse item i. For a dichotomous item, item information reaches its highest value at $\Theta = b$, which occurs when the probability of endorsing the ith item

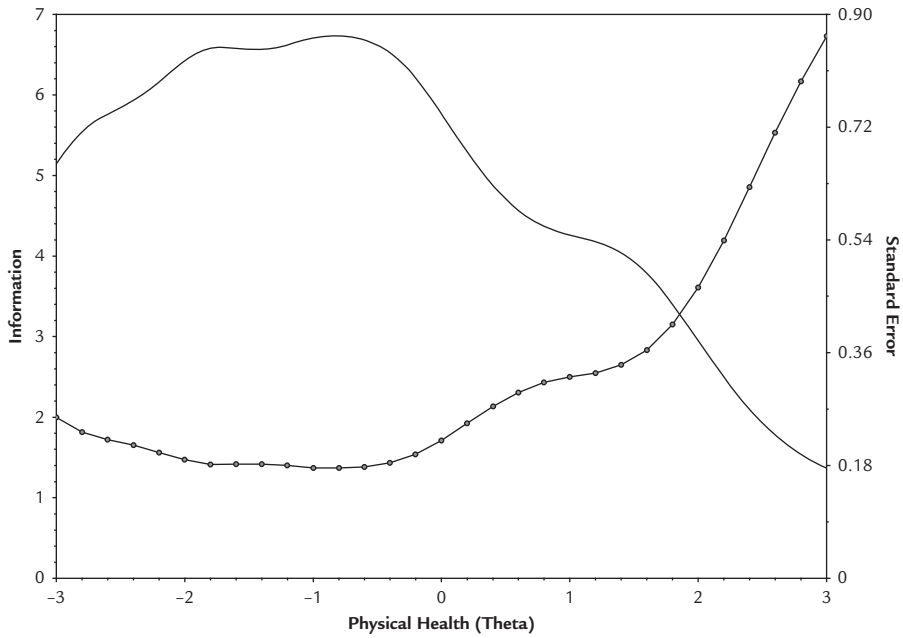

Figure 4. Scale information curve for PROMIS (Patient-Reported Outcomes Measurement Information System) 4-item Physical Health scale scores.

is $P_i = 0.5$. The amount of item information (precision) decreases as the item difficulty differs from the respondent's attribute level and is lowest at the extremes of the scale (ie, for those scoring very low or very high on the underlying concept).

Item information sums together to form scale information. Figure 4 shows scale information for the PROMIS Physical Health scale. Again, the *x*-axis shows the estimated Physical Health score (Θ) depicted on a *z*-score metric, with a more positive score representing better physical health. The left-hand side of the *y*-axis shows the information, and the right-hand side of the *y*-axis shows the SE of measurement. The peak of the curve shows where the Physical Health measure yields the greatest information about respondents (in the *z*-score range from –2 to –1). The SE of measurement is inversely related to, and a mirror image of, information (see subsequent discussion on the relationships between information, reliability, and the SE of measurement).

The item-information curve is peaked, providing more information and precision, when the *a* parameter (the item-discrimination parameter) is high; when the *a* parameter is low, the item-information curve is flat. If an item-information curve has a slope of $a = 1$, it has 4-fold the discriminating ability of an item with $a = 0.5$ (as seen by the squared term in the item-information function given in the preceding paragraph). The value of the *a* parameter can be negative; however, this results in a monotonically *decreasing* item response function. This implies that people with high amounts of the attribute have a *lower* probability of responding affirmatively in categories representing more of the concept than do people with lower amounts of the attribute. Such bad items should be weeded out of an item pool, especially if the parameter estimated has sufficient precision (ie, based on a sufficient sample size).

The reliability of measures scored on a *z*-score metric for the person-attribute parameter Θ (mean = 0 and SD = 1) is equal to $1 - SE^2$, where $SE = 1/(\text{information})^{1/2}$, and *SE* represents the SD associated with a given Θ. So information is directly related to reliability; for example, information of 10 is equivalent to reliability of 0.90. In addition to estimates of information for each item, IRT models yield information on the combination of items, such as a total scale score. Information typically varies by location along the underlying continuum of the attribute (ie, for

people who score low, in the middle, and high on the concept).

In Rasch measurement, the *person-separation index* is used as a reliability index because reliability reflects how accurately or precisely the scores separate or discriminate between persons; it is a summary of the genuine person separation relative to such separation and also measurement error.[9] *Measurement error* consists of both random error and systematic error and represents the discrepancy between scores obtained and their corresponding true scores. The person-separation index is based on the basic definition of *reliability* from classical test theory, as the ratio of true-score variance to observed variance (which equals the true-score variance plus the error variance). As noted earlier, the level of measurement error is not uniform across the range of a scale and is generally larger for more extreme (low and high) scores.

Person-Item Map

It is common to fix the mean of the item difficulties to equal 0. If the PRO measure is easy for the sample of persons, the mean across person attributes will be greater than zero ($\Theta > 0$); if the PRO measure is hard for the sample, the mean of Θ will be less than zero ($\Theta < 0$). Those most comfortable with the Rasch model (1-parameter model) produce *person-item* (or *Wright*) *maps* to show the relationship between item difficulty and person attribute. In principle, these maps can illuminate the extent of item coverage or comprehensiveness, the amount of redundancy, and the range of the attribute in the sample.

If the items have been written based on a construct map (a structured and ordered definition of the underlying attribute, as measured by the hierarchical positing of its series of items intended to be measured by the scale and conceived of in advance), the item map that follows the construct map can be used as evidence congruent with content validity.[16] A construct map is informed by a strong theory of which set of items require higher levels of the attribute for endorsement.

Figure 5 portrays such a person-item map of a 10-item scale on physical functioning.[9] Because of the scale content, the person attribute here is referred to as *person ability*. With a recall period of the preceding 4 weeks, each item is pegged to a different physical activity but raises the same question: "In the past 4 weeks, how difficult was it to perform the following activity?" Each item also has the same set of 5 response options: 1 = extremely difficult; 2 = very difficult; 3 = moderately difficult; 4 = slightly difficult; and 5 = not difficult. This example assumes that all activities were attempted by each respondent during the preceding 4-week recall interval.

Also assume that item difficulty (severity) emanated from the rating-scale model, a polytomous Rasch model, in which each item has its own difficulty parameter separate from the common set of categorical-threshold values across items.[17] If the more general partial-credit model were fit instead, the mean of the 4 categorical-threshold parameters for each item could be used to represent the difficulty of an item. If the response option were dichotomous instead of ordinal, a 1-parameter (Rasch) dichotomous logistic model could have been fit to obtain the set of item difficulties and attribute values.

At least 3 points are noteworthy. First, the questionnaire contains more easy items than hard ones, as 7 of the 10 items have location (logit) scores on item difficulty (severity) <0. Second, some items have the same difficulty scores, and not much scale information would be sacrificed if 1 of the dual items and 2 of the triplet items were removed. Third, patients tend to cluster at the higher end of the scale (note that the mean location score is ~ 1 for the ability of persons and that it exceeds the fixed mean location of 0 for difficulty of items), indicating that most of these patients would be likely to endorse (or respond favorably to) several of these items. Thus, either this group of patients had a high degree of physical functioning or, consistent with the previous evaluation of the items, there are not enough challenging or more difficult items, such as those on moderate activities (eg, moving a table, pushing a vacuum cleaner, bowling, or playing golf).

ICCs, which cannot intersect in the 1-parameter Rasch model, can intersect in the 2- and 3-parameter models if the discrimination parameters between items vary; such intersection can confound the item ordering.[9] In 2- and 3-parameter models, a consequence of this is that item 1 might be more difficult than item 2 for low levels of the attribute, whereas item 2 might be more difficult than item 1 for high levels of the attribute. In such cases, the item ordering will not correspond in the same way as the item-difficulty parameters. The Rasch model, which assumes equal item discrimination, corresponds exactly to (and is defined by) the order of item

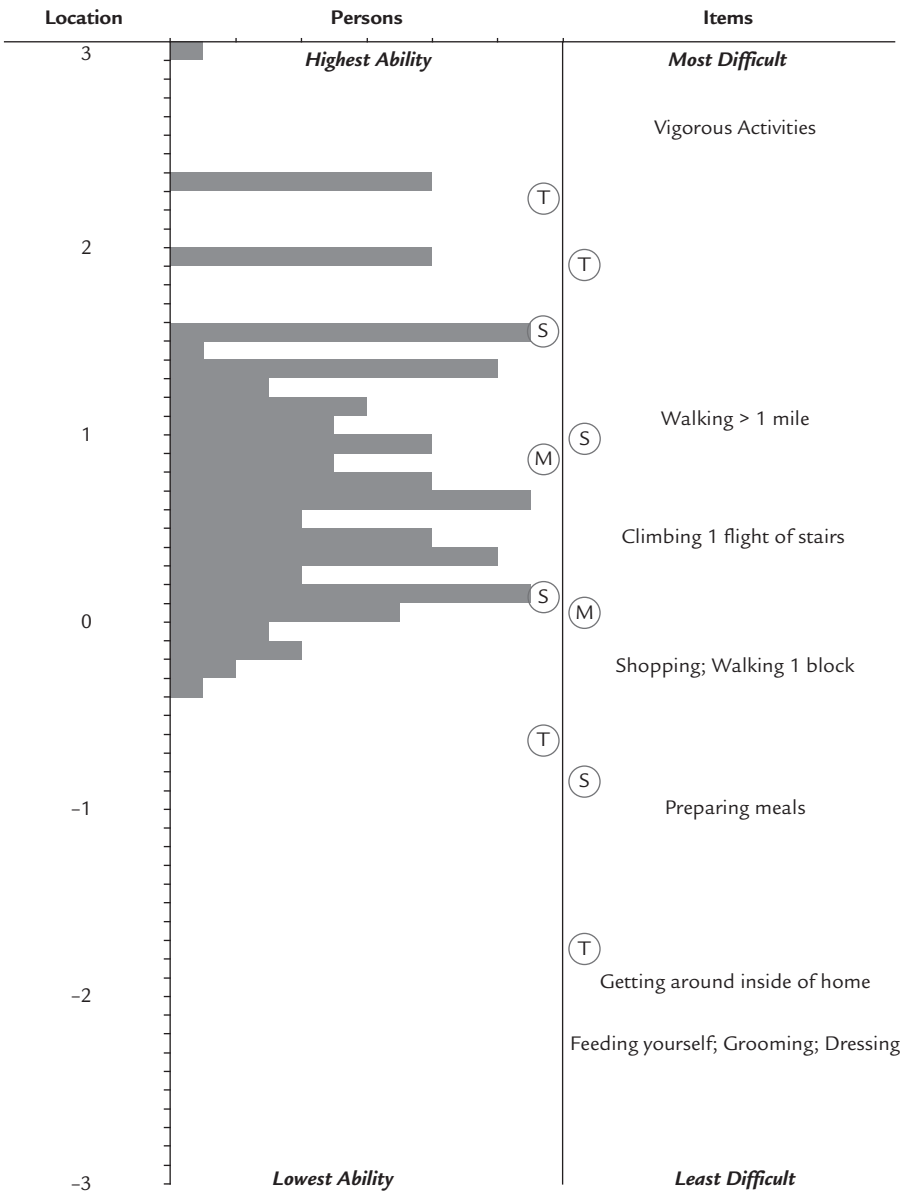

Figure 5. Illustration of person-item map on physical functioning. M = mean of person distribution or item distribution; S = 1 SD from the person mean or the item mean; T = 2 SD from the person mean or the item mean. Reprinted with permission.[9]

difficulties, and hence the order of the items is constant throughout levels of the attribute.

Assumptions in IRT

Before evaluating an IRT model, it is important to evaluate its underlying assumptions. Two assumptions (monotonicity and unidimensionality) are often used for scale evaluation in classical test theory, as noted previously. The assumption of *monotonicity*, which relates to the assumption of correct model specification, is met if the probability of endorsing each response category increases with the person's location on the attribute, and if the categories representing greater levels of the attribute require higher levels of the attribute to have a higher probability of being selected.

The assumption of *unidimensionality* of items in a scale is made so that a person's level on the underlying construct accounts fully for his or her responses to the items in the scale; it occurs when the correlation is absent or trivial between any pair of items for a fixed or given level of the attribute (also known as *local independence*). To satisfy this assumption, one can fit a factor-analytic model to the data to determine the extent to which there is sufficient unidimensionality. If the model fits the data well and there are no noteworthy residual correlations (ie, no such correlations ≥ 0.20), it provides support for the unidimensionality of the items in the scale.

Sample Size

IRT models, especially 2- and 3-parameter models, usually require large samples to obtain accurate and stable parameters; however, the 1-parameter (Rasch) model may be estimable with more moderate samples. Several factors are involved in sample-size estimation and no definitive answer can be given as to the sample size needed.[9]

First, the choice of IRT model affects the required sample size. One-parameter (Rasch) models involve the estimation of the fewest parameters and thus smaller sample sizes are needed, relative to 2- and 3-parameter models, to obtain stable parameter estimates of item difficulty and person location.

Second, the type of response options influences the required sample size. In general, as the number of response categories increases, a larger sample size is warranted because more item parameters must be estimated. It has been suggested that sample sizes of ≥ 200 are needed for the 1-parameter (Rasch) IRT model for dichotomous items.[18] At this sample size, SEs of item difficulty are in the range of 0.14 to 0.21 (based on [2/(square root of n)] $<$ SE $<$ [3/(square root of n)], where n is the sample size).[19] Another suggestion is that, for item-difficulty (and person-measure) calibration to be within 1 logit of a stable value with 95% confidence, an expected sample size as small as 30 subjects would suffice in a Rasch model for dichotomous items (a larger sample size is needed for polytomous items).[20] To be within 1 logit of a stable value for a dichotomous item targeted to have probability of endorsement of 50% means that the true probability of endorsing the item can be as low as 27% and as high as 73%—a wide range. The challenge with Rasch analysis is that, because it requires model-based estimation, it needs a sizeable sample to yield stable results. This reason has led some researchers to conclude that results of Rasch analyses of data from small sample sizes have the potential to be misleading and therefore small sample sizes are not recommended.[21] In 2-parameter (eg, graded-response) models, a sample size of ≥ 500 is recommended.[18] Although ≥ 500 is ideal, a much smaller sample could still provide useful information, depending on the properties and composition of the scale. In general, the ideal situation is to have adequate representation of respondents for each combination of all possible response patterns across a set of items—something that is rarely achieved. It is important, though, to have at least some people respond to each of the categories of every item to allow the IRT model to be fully estimated.

Third, study purpose can affect the necessary sample size. A large sample size is not needed to obtain a clear and unambiguous picture of response behavior or trends, and therefore a large sample size is not generally needed in the instrument-development stage to demonstrate content validity, provided that a heterogeneous sample is obtained that accurately reflects the range of population diversity inherent in item and person responses. If the purpose is to obtain precise measurements of item characteristics and person scores, with stable item and person calibrations, sample sizes in the hundreds are generally required; one recommendation is sample sizes > 500.[22]

Fourth, the sample distribution of respondents is an important consideration. Ideally, respondents should be distributed fairly uniformly over the range of the attribute (construct) being measured. If fewer people are located at the ends of the attribute, items also positioned at the extreme ends of the construct will have higher SEs associated with their parameters.

Fifth, measures with more items may require larger sample sizes. Additional items increase the possibility that the parameters of any 1 item need a larger sample to be adequately estimated.

Finally, if the set of items in a questionnaire has a poor, or merely a modest, relationship with the attribute (which is not unexpected during the content-validity stage of instrument development), a larger sample size would be needed because more information is needed to compensate for the smaller size of the relationship. If the relationship of items with the attribute is small,

however, the credibility of the scale or at least some of its items should be called into question.

DISCUSSION

Classical test theory and IRT provide useful methods for assessing content validity during the early development of a PRO measure. IRT requires several items so that there is adequate opportunity to have a sufficient range for levels of item difficulty and person attribute. Single-item measures, or too few items, are not suitable for IRT analysis (or, for that matter, for some analyses in classical test theory). In IRT and classical test theory, each item should be distinct from the others, yet should be similar and consistent with them in reflecting all important respects of the underlying attribute or construct.[23] For example, a high level of cognitive ability (the attribute of interest) implies that the person also has high levels of items constituting cognitive ability, such as vocabulary, problem solving, and mathematical skills.

IRT can provide information above and beyond classical test theory, but estimates from IRT models require an adequate sample size. Sample-size considerations for IRT models are not straightforward and depend on several factors, such as the numbers of items and response categories.[9] Small sample sizes should be discouraged when fitting IRT models because their model-based algorithms are not suited for small samples. Rasch models involve the estimation of the fewest parameters and thus smaller sample sizes are needed, relative to 2-parameter models, to obtain stable parameter estimates on item difficulty and person location. That said, even Rasch models require a large enough sample to achieve reliable results for deciding what items to include and what response categories to revise.[21]

If a researcher has few qualitative data and wants to get preliminary information about the content validity of the instrument, then descriptive assessments using classical test theory should be the first step. As the sample size grows during subsequent stages of instrument development, confidence in the numerical estimates from Rasch and other IRT models (as well as those of classical test theory) would also grow. In later stages of PRO development, researchers could strive for a sample of, say, 500 individuals for full psychometric testing. If the construct of interest is well-defined and responses are sufficiently dispersed along the trait continuum, significantly smaller sample sizes may be sufficient.

A Rasch model may be more amenable for the developmental stages of PRO measures than other IRT models, because of its item- and person-fit indices, person-item map, and smaller sample-size requirements. Compared with classical test theory, a Rasch model (and other IRT models) provides the distinct benefit of a person-item map. The visual appeal of this map enriches understanding and interpretation in suggesting the extent to which the items cover the targeted range of the underlying scale and whether the items align with the target patient population.

CONCLUSIONS

The present article presents an overview of classical test theory and IRT in the quantitative assessment of items and scales during the content-validity phase of PRO-measure development. Depending on the particular type of measure and the specific circumstances, either approach or both approaches may be useful to help maximize the content validity of a PRO measure.

ACKNOWLEDGMENTS

The authors gratefully acknowledge comments from Dr. Stephen Coons (Critical Path Institute) on the manuscript and also the comprehensive set of comments from 2 anonymous reviewers, all of which improved the quality of the article.

Dr. Lundy is an employee of the Critical Path Institute, which is supported by grant No. U01FD 003865 from the United States Food and Drug Administration. Dr. Hays was supported in part by funding from the Critical Path Institute and by grants from the Agency for Healthcare Research and Quality (2U18 HS016980), the National Institute on Aging (P30AG021684), and the National Institute of Minority Health and Health Disparities (2P20MD000182).

Each author made a substantial contribution to the conception, design, and content of the manuscript; was involved in drafting the manuscript and revising it critically for important intellectual content; has given final approval of the version to be published; and has agreed to be accountable for all aspects of the work.

CONFLICTS OF INTEREST

J. Cappelleri is an employee of, and holds stock options in, Pfizer Inc. The opinions expressed here

do not reflect the views of Pfizer Inc or any other institution. The authors have indicated that they have no other conflicts of interest with regard to the content of this article.

REFERENCES

1. US Dept of Health and Human Services, Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. *Fed Reg*. 2009;74:65132–65133.
2. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1—eliciting concepts for a new PRO instrument. *Value Health*. 2011;14:967–977.
3. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 2—assessing respondent understanding. *Value Health*. 2011;14:978–988.
4. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. *Ann Rev Clin Psychol*. 2009;5:1–25.
5. Anastasi A, Urbina S. Psychological Testing. 7th ed. Upper Saddle River, NJ: Prentice Hall; 1997.
6. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD*. 2005;2:63–67.
7. Hays RD, Wang E. Multitrait scaling program: MULTI. In: *Proceedings of the Seventeenth Annual SAS Users Group International Conference*. 1992:1151–1156.
8. Hays RD, Fayers P. Evaluating multi-item scales. In: Fayers P, Hays RD, eds. *Assessing Quality of Life in Clinical Trials: Methods and Practice*. 2nd ed. Oxford, UK: Oxford University Press; 2005:41–53.
9. Cappelleri JC, Zou KH, Bushmakin AG, et al. Patient-Reported Outcomes: Measurements, Implementation and Interpretation. Boca Raton, Fla: Chapman & Hall/CRC Press; 2014.
10. Thissen D, Wainer H. Some standard errors in item response theory. *Psychometrika*. 1982;47:397–412.
11. Tabachnick BG, Fidell LS. Using Multivariate Statistics. 3rd ed. New York, NY: Harper Collins; 1996.
12. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. *Psychol Methods*. 1999;4:84–99.
13. Brown TA. Confirmatory Factor Analysis for Applied Research. New York, NY: Guilford; 2006.
14. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care*. 2000;38(Suppl): II-28–II-42.
15. Hays RD, Bjorner JB, Revicki DA, et al. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. *Qual Life Res*. 2009;18:873–880.
16. Wilson M. Constructing Measures: An Item Response Modeling Approach. Mahwah, NJ: Lawrence Erlbaum Associates; 2005.
17. Bond TG, Box CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.
18. Reeve B, Fayers P. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, eds. *Assessing Quality of Life in Clinical Trials*. 2nd ed. New York, NY: Oxford University Press; 2005:53–73.
19. Wright BD, Stone MH. Best Test Design. Chicago, Ill: MESA Press; 1979.
20. Linacre M. *Sample size and item calibration (or person measure) stability*. 2010. http://www.rasch.org/rmt/rmt74m.htm. Accessed December 1, 2013.
21. Chen WH, Lenderking W, Jin Y, et al. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behaviour item bank data. *Qual Life Res*. 2014;23:485–493.
22. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
23. Fayers FM, Machin D. Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes. 2nd ed. Chichester, UK: John Wiley & Sons; 2007.

Address correspondence to: Joseph C. Cappelleri, PhD, 445 Eastern Point Road MS 8260-2502, Groton, CT 06340. E-mail: joseph.c.cappelleri@pfizer.com