What Is Sufficient Evidence for the Reliability and Validity of Patient-Reported Outcome Measures?

Marlene H. Frost, RN, PhD,1 Bryce B. Reeve, PhD,2 Astra M. Liepa, PharmD,3 Joseph W. Stauffer, DO,4 Ron D. Hays, PhD,5 the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group

1Women’s Cancer Program, Mayo Clinic, Rochester, MN, USA; 2Outcomes Research Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA; 3Global Health Outcomes, Eli Lilly and Company, Indianapolis, IN, USA; 4Global Medical Affairs, Alpharma, Piscataway, NJ, USA; 5UCLA Division of General Internal Medicine and Health Services Research, Department of Medicine, UCLA Department of Health Sciences, Los Angeles, CA, USA; *RAND Corporation, Santa Monica, CA, USA

ABSTRACT

This article focuses on the necessary psychometric properties of a patient-reported outcomes (PROs) measure. Topics include the importance of reliability and validity, psychometric approaches used to provide reliability and validity estimates, the kinds of evidence needed to indicate that a PRO has a sufficient level of reliability and validity, contexts that may affect psychometric properties, methods available to evaluate PRO instruments when the context varies, and types of reliability and validity testing that are appropriate during different phases of clinical trials.

Points discussed include the perspective that the psychometric properties of reliability and validity are on a continuum in which the more evidence one has, the greater confidence there is in the value of the PRO data. Construct validity is the type of validity most frequently used with PRO instruments as few “gold standards” exist to allow the use of criterion validity and content validity by itself only provides beginning evidence of validity.

Several guidelines are recommended for establishing sufficient evidence of reliability and validity. For clinical trials, a minimum reliability threshold of 0.70 is recommended. Sample sizes for testing should include at least 200 cases and results should be replicated in at least one additional sample. At least one full report on the development of the instrument and one on the use of the instrument are deemed necessary to evaluate the PRO psychometric properties. Psychometric testing ideally occurs before the initiation of Phase III trials. When testing does not occur prior to a Phase III trial, considerable risk is posed in relation to the ability to substantiate the use of the PRO data. Various qualitative (e.g., focus groups, behavioral coding, cognitive interviews) and quantitative approaches (e.g., differential item functioning testing) are useful in evaluating the reliability and validity of PRO instruments.

Keywords: patient-reported outcomes, psychometric, validation.

Introduction

Evaluation of health-care interventions in clinical trials has traditionally focused on efficacy and safety. Patient-reported outcomes (PROs), outcomes directly reported by the patient, are increasingly being considered and used as part of these evaluations. PROs include symptoms and other aspects of health-related quality of life such as physical or social function, treatment adherence, and satisfaction with treatment [1,2]. As with more traditional efficacy and safety end points, reliable and valid instruments are needed to measure PROs.

PROs are distinct from traditional clinical efficacy measures (e.g., survival in cancer, smoking cessation rates, laboratory test results). PROs directly reflect the impact of disease and its treatment from the patient’s perspective and can measure the tradeoff between efficacy of the treatment and what the patient is willing to tolerate [3]. PROs are especially significant when symptoms, functioning, and well-being are important outcomes or areas of concern. They are particularly informative when interventions reveal otherwise similar efficacy and safety using traditional clinical measures or when an intervention provides only a small clinical benefit. Because PROs reflect the patient’s perspective, they have the potential to facilitate patient involvement in treatment decision-making and provide guidance for health-care decisions.

The ability of a PRO measure to improve decision-making in clinical research relies on the psychometric strength of the instrument to capture the burden of disease or treatment. Reliability and validity are essential psychometric properties for any measure [4]. Evidence for an instrument’s reliability and validity falls
along a continuum from no evaluation to full evaluation for the study population. Thus, both reliability and validity are more accurately described as “continuous” rather than “dichotomous” psychometric indices. For this reason, claiming that an instrument is completely “reliable” or “valid” is inaccurate. Similarly, saying an instrument has been “validated” conveys no information other than to say its performance or psychometric properties have been evaluated. The more evidence that the instrument is reliably measuring the specific PRO it is supposed to be measuring the more confidence one has in it.

Reliability and validity are separate psychometric properties. An instrument that is not reliable (internal consistency, test–retest) by definition cannot be valid. Measures can be highly reliable but not measure what they are purported to measure. For example, a self-report of year in school (e.g., third, fourth, fifth grade) by elementary students may be very reliable but this would not be a valid measure of grade school success because of the low dropout rate at that age and large variation in performance within the same grade level. Thus, reliability is necessary but not sufficient in determining validity.

This article focuses on the necessary psychometric properties to support use of PROs in health-care intervention evaluations in clinical trials. Specifically addressed are: 1) the importance of reliability and validity; 2) the psychometric approaches used to provide reliability and validity estimates; 3) the evidence needed to indicate that a PRO has a sufficient level of reliability and validity; 4) contexts that may affect psychometric properties; 5) methods available to establish the reliability and validity of PRO instruments when the context varies; and 6) types of reliability and validity testing that are appropriate during different phases of clinical trials. The discussions are intended to address the need to evaluate instrument properties as proposed by the Food and Drug Administration (FDA) in its guidance on PROs for labeling claims (see Fig. 1) [5].

**Reliability and Validity Performance Recommendations**

**Reliability**

The first attribute evaluated and reported typically is reliability, the extent to which a measure yields the same number or score each time it is administered when the construct being measured has not changed. Internal consistency reliability, the primary method of estimating reliability for multi-item scales, provides information about the associations among different items in the scale. Internal consistency is typically indexed by Cronbach’s coefficient alpha, which is estimated using a two-way fixed-effect analysis of variance (ANOVA) that partitions the “signal” (i.e., between person variance) from the “noise” (i.e., interaction between people and responses to different items). Alpha can also be expressed using the formula:

\[ \text{Alpha} = \frac{(K \cdot \text{R}_{ii})}{(1 + (K - 1) \cdot \text{R}_{ii})} \]

This alternative expression illustrates how reliability increases with the number of items (K) in a scale and the strength of the correlations among items as represented by the intraclass correlation (\( \text{R}_{ii} \)). \( \text{R}_{ii} \) represents the estimated reliability for a single item. Applying the formula, a scale with an intraclass correlation of 0.30 and five items will have an estimated reliability of 0.68. Thus, a PRO measurement with multi-item scales yields more precise measurement of PRO constructs than a single-item measure.
Beyond statistical considerations, multi-item scales better capture multiple attributes that underlie PRO domains like pain, fatigue, or nausea. Although the trade-off with longer scales is increased response burden, the improved precision and validity added with well designed multi-item scales (over single-item measures) enhances confidence in any decision regarding clinical effectiveness. Coefficient alpha would not be appropriate for single item scales or to compute coefficient alpha for a set of items that measure different constructs (e.g., a variety of potentially unrelated symptoms).

Picking the optimal time interval for test–retest reliability may be difficult. Repeat administration timing should not be so soon that responses at the second assessment are simply memories of the first assessment; yet it should also not be so long that true change in the construct has occurred during the time interval between the initial and subsequent assessment.

Reliability coefficients range in theory between 0 and 1, with 0.70 the standard threshold for adequate reliability for use of measures for group comparisons. For individual applications, a more stringent minimum threshold of 0.90 reliability has been suggested [4,6]. The higher standard is needed because the error around an individual’s score is larger than the error around a group mean score. For example, even with a reliability of 0.91, the individual’s standard error of measurement (SEM) (standard deviation [SD] * square root of 1—reliability coefficient) is equal to 0.30 * SD. If the SD of a measure is 10 as it is with the SF-36v2™ scales, then the width of the 95% confidence interval (CI) around an individual’s score is 12 points (greater than a SD) as it extends from 6 points below to 6 points above the estimated true score. Using the same instrument in a group, a sample size of 25 people would yield a width of the 95% CI around the group mean of 2.4 points (approximately one-quarter of a SD). Clinical trials are group applications and because the error of measurement of group means is driven primarily by sample size, a 0.70 reliability threshold is appropriate.

Reliability and SEM are inversely related. This association has important implications for the sample size needed to detect group differences (longitudinally or cross-sectionally) in PRO measures. For example, adjusting for the SEM, the required sample size per group to detect one-third of a SD difference between baseline and follow-up is 297 versus 227 per group if the reliability is 0.69 versus 0.84, respectively [7,8]. Thus, increased reliability in the PRO measure reduces the sample size needed in the example by 70 people.

The limitation of the traditional measures of reliability is that they assume that the reliability of a scale is fixed for all score levels. For example, a pain instrument with a reliability of 0.82 would be considered acceptable for measuring a group’s average state of pain no matter if the group experiences mild, average, or severe levels of pain.

In contrast, item response theory (IRT) provides an assessment of reliability of item and scale information curves. The IRT information curve indicates the precision of an item or scale for measuring different levels along its underlying trait continuum (see Fig. 2). Items are most useful when they are appropriate or matched to the individual completing it. For example, asking a person who is generally happy and content with life about thoughts of suicide in the last week is not likely to be informative for measuring his or her emotional distress level. Items are most informative when the answer that someone will give is less certain (e.g., just as likely to say “yes” as “no” to a dichotomous question). Because of the emphasis on assessing dysfunction, information curves for PRO measures often reflect higher precision for measuring low function than for measuring high function.

**Item Discrimination across Scales: Approach to Enhance Reliability and Validity**

Internal consistency reliability provides information about the associations among different items in the scale. Equally important is to evaluate the extent to which items in one scale distinctively represent that construct rather than any other related constructs. One approach to evaluating the distinctiveness of different measures is multi-trait scaling analysis [9]. This method examines correlations between items and scales, which indicates if the item uniquely represents the hypothesized scale. Confirmatory factor analysis provides a parallel and more sophisticated approach to evaluating item discrimination across scales [10].

**Validity**

A PRO measure must distinctly define one construct and not overlap with other putatively distinct concepts. Validity refers to the extent to which an instrument measures what it was intended to measure and not something else. Generally, a strong correlation should be demonstrated with measures addressing similar constructs and a weak correlation with measures addressing disparate constructs. There are three main subtypes of validity: content, criterion, and construct.

**Content validity.** This is the extent to which the PRO instrument measures the appropriate content and represents the variety of attributes that make up the measured construct. Another way of expressing content validity is the adequacy of sampling of the material in the measure, which is best ensured by a plan for content and item construction before the measure is developed. Focus groups and other qualitative methods (e.g., cognitive interviews) are sources for
appropriate content. A group of experts can examine items and can either endorse the content validity or identify any important gaps in content. Face validity assesses the content of a scale in terms of the extent to which it is perceived to be measuring what it is supposed to measure by patients and experts.

**Criterion validity.** This refers to the extent to which the measure agrees with an external standard measure. An example would be the development of an observational measure of how well an individual is breathing and comparing it to a pulse oximeter reading that measures oxygen saturation.

Because PRO measures typically have no standard, criterion validity is usually not applicable. For situations where it is appropriate, evaluation of criterion validity would involve determining the extent to which the new measure is consistent or captures the essence of the standard measure. For example, one might employ contingency table analyses of sensitivity and specificity or area under the curve analyses to assess the level of agreement of the new measure with the standard [11].

**Construct validity.** This is the extent to which the measure “behaves” in a way consistent with theoretical hypotheses; it represents how well scores on the instrument are indicative of the theoretical PRO construct. Construct validity evaluation includes the degree to which a measure correlates with other measures to which it is similar and does not correlate with
(diverges from) measures that are dissimilar. A surplus of terminology exists in the literature that falls into the general class of construct validity. For example, the multi-trait-multi-method of validity assessment refers to convergent and discriminant validity as aspects of construct validity.

Responsiveness measures an instrument’s ability to capture change [12]. Responsiveness is often overlooked and yet is vital when using repeated measures over time when the concept is known to have changed.

Construct validity is typically examined using bivariate correlations, factor analysis, and multivariate regression models. For example, one could hypothesize that a breast cancer patient’s self-esteem was positively associated with breast-conserving surgery. One could regress self-esteem on type of surgery and background variables such as age, marital status, and educational level. A statistically significant finding for type of surgery would support the hypothesis and construct validity of the measure of self-esteem.

Level of Psychometric Evidence to Support Use of PROs in Clinical Trials

Both qualitative and quantitative information are important in evaluating the level of evidence to support reliability and validity. Multiple pieces of evidence should support use of a PRO measure (e.g., results of two or more focus groups for content validity, strong correlations with two similar measures for construct validity). Qualitative data require investigators to make a solid case about saturation of information (i.e., documentation that further studies do not generate new information).

Sample sizes for quantitative analyses should be large enough to meet a desired level of measurement precision or standard error [13]. With sample sizes of 100, 200, 300, and 400, the standard errors around a correlation are approximately 0.05, 0.03, 0.03, and 0.02, respectively. A sample of 300 yields 84% power (alpha = 0.05, two-tailed) to detect a correlation between two variables of 0.17. For factor analyses, at least five cases per variable and a minimum of 300 cases have been recommended [14] and Charter [15] argued for a minimum of 400 subjects for reliability and validity studies. Reeve and Fayers [16] suggested sample sizes of at least 200 for the simple one-parameter (e.g., Rasch) IRT model. At this sample size, the modeled item standard errors (2/(square root [N]) < standard error < 3/(square root[N])) are in the range of 0.14–0.21 [17]. For two-parameter (e.g., graded response) models, a sample size of at least 500 was recommended [16]. Hence, the sample size requirements for reliability and validity assessment depend on the specific circumstances and analytical tools, but a minimum of about 200 cases is suggested for even the most basic psychometric analyses. If a measure is to be used in a specific subgroup (e.g., African Americans), then sufficient sample size is needed to represent that subgroup. In some situations (e.g., when large patient accrual is not feasible) a smaller sample size might be considered sufficient. In this situation, analytical methods can include simple descriptive statistics of correlations among items and subscales.

Replication of psychometric estimates is needed either by a sufficiently large and representative sample that can be split into two subsamples for cross-validation or two samples of sufficient sample size. One sample is used to explore the properties of the scale and the second sample is used to confirm findings found with the first sample. If the results of the two samples are inconsistent, then psychometric estimates from another sample may be required to establish the properties of the measure. Clinical trial data can be used to support the psychometric properties of an instrument.

When evaluating the reliability and validity of an existing instrument, a review of the literature typically reveals the needed information. Nevertheless, a more comprehensive evaluation may also involve sponsor communication with the developer or other investigators to identify unpublished data. The goal of the evaluation is to determine the level of evidence that is available in circumstances (e.g., population, mode of administration, intervention) similar to those of the proposed clinical trial. The amount of existing documentation may depend on how long the instrument has been available, the population(s) in which it has been used, prevalence of the conditions measured, and the specificity of the instrument to certain subgroups. When the existing evidence is not sufficient, additional documentation must be produced.

Investigators should generate at least one full report on the development and evaluation of the instrument. In addition, they should have at least one report on the use of the instrument in an interventional setting, such as a clinical trial or naturalistic study, to confirm feasibility and psychometric attributes. These reports should provide detailed information on patient population, sample size, instrument administration, scoring methods, statistical analyses, and evidence of stability of findings.

When PROs are secondary end points in clinical trials, the level of detail provided in the primary publication of the results of the trial rarely includes the psychometric properties of the PRO instrument. Additionally, there may be no secondary publications that focus on the PRO end point. Thus, to obtain the PRO psychometric properties, follow-up with the developer or investigators may be indicated. Availability of reports may be limited if publications are still under review or if the developer or investigator prefers to restrict use of the PRO measure for proprietary reasons.
**Contexts That May Affect Psychometric Properties**

Although information about the reliability and validity of an instrument is critical, these properties must be considered in the context of the setting in which the instrument will be used. The context may change the ability of the instrument to measure the desired construct. The following section outlines some factors that should be considered when evaluating available reports.

**Patient Population**

Particular attention should be given to the comparability of patient populations in reports relative to the target population. Unique characteristics of the target population that might result in variations in the results of a PRO assessment should be identified in advance. These might include age, gender, race and ethnicity, marital status, socioeconomic status, education, comorbidities, disease status whether the condition is acute or chronic, and whether the intervention is curative or palliative.

A PRO item bank might be used to address a variety of contextual concerns, such as population differences and repeated measures, which may affect reliability and validity. An item bank is a collection of items that: 1) measure a single domain; 2) have undergone rigorous qualitative, cognitive and psychometric review (including cross-cultural group validations); and 3) have been IRT-calibrated with a set of properties allowing instrument developers to select an item set matched to the characteristics of the study population [18].

Investigators can use the bank to reduce response burden by developing short-form instruments that select subset of items. Turner et al. (2007) in this supplement review the value and potential of item banks (article 3 of this series) [19]. The strength of a bank to deliver reliable, valid, and efficient measurement depends, as with any other PRO instrument, on the developmental process. Similar to the developmental process outlined in Fig. 1, an item bank should start with a conceptual framework that leads to identifying existing items that measure that domain and/or developing new items (see Rothman et al. in this supplement) [20]. Once the item pool has been built of old and new items, users must carry out a thorough qualitative review phase beginning with evaluating the items. Next, response data with a large sample representative of the target population must be collected and used to review the item performance quantitatively and make IRT-calibrations. The quantitative review phase requires assessing the psychometric domains of reliability and validity described in this and other articles [21–23]. Essentially, the attributes that make up a quality item bank do not differ materially from the attributes of any other PRO instrument under consideration.

**Cross-cultural application.** Patient-reported outcomes instrument development and initial application may be from a single institution; this factor frequently limits the generalization to other settings and patient populations, including cultural backgrounds. Ideally, additional PRO measure reports will have been applied across multiple study sites and potentially across multiple countries to address cross-cultural issues. Documentation that appropriate cross-cultural validation methods were used in the development of translations should also be available [24].

**Recall**

A critical issue is the instrument’s recall (or reference) period. The recall period should be considered relative to the disease condition and the intervention. Symptoms that change every day, such as level of pain or fatigue, pose difficulties when individuals are asked to recall this information. Nevertheless, with other areas in which changes may not take place every day or even every week, recall over the last month may be more informative. Examination of the bias associated with recall can be addressed by ecologic momentary assessment [25]. With this approach, investigators survey individuals in environments that represent their real world, they collect data at the moment the situation of interest is occurring, and they take multiple samplings. In general, recall over a few days to a few weeks is less problematic than recall over a longer period of time.

**Variations in Instrument Administration**

**Mode of administration.** The method of data collection (e.g., in-person interview, article, telephone, electronic, Internet, proxy) can produce different systematic results, raising questions about the interchangeability of the methods [26]. To evaluate mode effects, a substudy can be conducted in which study participants are randomized to different modes of administration. Item missing data, means, SDs, item-scale correlations, reliability, and correlations of items and scales with other variables can be compared by mode.

**Timing of assessments.** Frequency or timing of assessments may alter findings in situations where a patient’s condition is unstable because of disease or an intervention, if an intervention causes acute toxicity such as with cyclic administration of cytotoxic chemotherapy, or if surgical interventions have occurred.

**Proxy measures.** Investigators need to be cautious in trying substitute proxy reports for patient self-report data because of inherent differences between the two
types of respondents. The more observable the function, the greater the agreement between a proxy and a patient’s report; better concordance between proxies and patients tends to occur for the physical domains than for psychosocial domains [27,28]. Proxies tend to report more disability and depression about the patients than patients report about themselves. In contrast, proxies tend to attribute higher levels of cognitive ability to the patients than do patients when rating their own cognitive ability. For specific cognitive and depression instruments such as the Mini Mental State Examination and the Center for Epidemiologic Studies depression scale, scoring on a continuous scale compared to an impaired/unimpaired dichotomy showed a higher level of concordance between patient and proxy than did an impaired/unimpaired dichotomy. Nevertheless, an impaired/unimpaired dichotomy, compared to a continuous scale, results in less of a biased estimate of impairment [27].

**Statistical Analysis**

Investigators need to determine the reasons for the presence or absence of statistically or clinically significant differences between groups. A lack of differences may be a result of the PRO instrument lacking the sensitivity needed (articles Sloan et al. [29] and Snyder et al. [30] in this series cover these points in more detail). Nevertheless, this may also be a result of limited sample size (see sample size discussion in Level of Psychometric Evidence to Support Use of PROs in Clinical Trials section), missing data, or appropriateness of analysis.

The apparent sensitivity of the PRO measure may depend on how the data are reported. For example, broad PRO claims such as overall health-related quality of life and total symptom scores can obscure the changes that are occurring within individual concepts or symptoms. A summary score may obscure the fact that the person is having great difficulty with social function when the person is doing well on the other aspects of health-related quality of life. The total score does not reflect this specific difficulty. Thus, looking at specific domains as a profile along with the summary score(s) is critical [31]. Additionally, reporting only group differences may obscure individual patient changes that are significant.

**Stability of Findings**

When considering a PRO instrument that does not appear to produce consistent findings, investigators need to determine whether this lack of consistency occurred because of differences in methodology of the reports including variations in number of investigative sites, population characteristics, sample size, instrument administration, analysis, and the intervention. If any estimates fall outside of accepted values, this variation should be explained. This explanation is most critical if this variation is reported in a setting similar to the target indication or population. A compelling case would need to be built as to why this particular PRO measure would be used in light of inconsistent reports.

**Methods to Evaluate PRO Instruments When the Format and Application of the PRO Is Different from Previous Psychometric Evaluation Studies**

Determining the need for additional psychometric evaluation of a PRO measure is tricky. Some suggest (in the context of using an instrument in a new cultural population) any small changes in the items, formulation, or design can affect respondents’ understanding of the items or accuracy of the measurement [32]. Ideally, an instrument would be reevaluated when any changes or adaptations occur, but burden, time, and budgetary constraints may make further psychometric testing infeasible. The article by Snyder et al. [30] in this supplement addresses these issues and makes recommendations under what conditions sponsors should feel comfortable that the properties of the instrument are robust to minor modifications. Nevertheless, if a sponsor decides to reevaluate the instrument, then several qualitative and quantitative approaches or tools are available (Fig. 3); we discuss their strengths and limits next. The same approaches or tools can be used to initially create an instrument as presented by Turner et al. in this series [19].

**Qualitative Approaches**

**Focus groups.** Focus groups typically consist of a moderator interacting with 6 to 12 people representing the target population, or significant others (e.g., family, health-care providers) involved with the target population. Focus groups may be used during all stages of instrument development and evaluation (see Fig. 3). In early stages, focus groups may respond to open-ended questions that elicit information about important issues and concerns about the PRO construct. This may uncover cultural differences in the experiences of the PRO domain. Further, focus groups can produce feedback on item formulation and how items are perceived. For lengthy instruments, focus group members typically complete the instrument in advance and the moderator may ask individuals to discuss complex terms and identify unclear items. Focus groups may also help in generating hypotheses or explanations for interpreting data that have been collected [33].

**Behavioral coding.** In behavioral coding, the interactions between the interviewer and respondent in interviews are recorded, coded, and analyzed to identify overt sources of error and to evaluate the adequacy of
proposed questions. This process can be labor intensive, especially if the researchers code all aspects of the interview including response time. Alternate approaches may rely on a limited set of codes that represent key behaviors; examples include respondent need for an item repeated or clarified [33]. For assessing conceptual equivalence across racial or cultural groups, computer-assisted telephone interviews or computer-assisted personal interviews are conducted for each group. Errors associated with one group can identify problems in the translation or cultural relevance of the items. Behavioral coding provides aggregate coding summaries of interviews conducted in samples of 50 respondents or more [34,35] although sample sizes significantly vary across studies.

Cognitive interviewing. Cognitive interviewing is a powerful tool for gaining a better understanding of the underlying or covert process involved in responding to survey items through the use of verbal probing techniques [36]. It is used to evaluate the quality of each item in terms of a person’s understanding of the item, ability to retrieve the appropriate information, decision-making on reporting retrieved information, and selection of the response. Further, cognitive interviews can be used to examine relationships between participant characteristics, such as ethnicity, and responses to survey items.

The cognitive interview process includes both the administration of an instrument and the collection of additional verbal information about the survey responses [36]. Cognitive interviewing encompasses the more specific practice of cognitive debriefing. The term “cognitive debriefing” is typically associated with following-up with a respondent through the use of retrospective probes after they have completed an instrument with a line of questions aimed at uncovering any difficulties the person may have experienced with either the item content or instructions. For cognitive interviews that involve concurrent probing, the interviewer follows each question with a series of probes to capture patient understanding. In contrast to the retrospective cognitive debriefing, concurrent probing can yield information about the cognitive processing of the item at a point close in time to when it is first presented. Using 5–12 persons is recommended for cognitive interviews, with a second, iterative round of testing to evaluate items revised from the first round [35].

Cognitive interviews can include two types of probes. Scripted probes ensure that particular information is collected in every interview so that it can be compared across all interviews. Emergent, nonscripted probes to help interviewers make sense of gaps or contradictions in participants’ responses and provide contextual information needed to define item problems precisely. If sufficient numbers of cognitive interviews are conducted, a coding mechanism allows researchers to utilize quantitative methods (e.g., logistic regression) to determine if problems encountered during the interviews can be attributed to specific factors, such as cultural effects. Cognitive interviewing has been employed as an instrument evaluation tool in several PRO studies [32,36,37].

Quantitative Approaches

Differential item functioning. Several quantitative approaches are available for evaluating measurement invariance in the form of differential item functioning (DIF) testing. DIF tests whether one group responds differently to an item than another group despite controlling for differences between the two groups on the
measured construct. Between-group comparisons for instruments containing such items are problematic, because scores from the two groups may be indicative of a variety of factors other than those the instrument is intended to measure.

Methods for DIF testing include IRT modeling [37–39], structural equation modeling [40], and regression models [41]. DIF methods typically examine whether the associations between the estimated underlying attribute and items in the instrument vary by subgroup. DIF testing is performed on an item-by-item basis controlling for group differences using the other items within the instrument or covariates. Any empirical finding that DIF is present should be followed by the other qualitative techniques to understand the underlying causes.

Each qualitative and quantitative method above has its strengths and weaknesses in terms of gained information, resource constraints, and burden on the administrators and respondents [35]. Each can be applied at different stages of instrument development and evaluation. Incorporating more than one methodology enhances the strength of the validation process (see Fig. 3). When problematic items have been identified and the source of the problem understood, items might be rewritten, removed from the instrument, or statistically controlled for when computing individual scores.

**Additional Sources to Enhance Validity and Complement the PRO**

Sources for validating PRO instruments or increasing the information gained from PROs can be found in other clinical-based or performance-based measures. For instance, a surgical patient’s hemoglobin level provides information relevant to the patient’s subjective report of fatigue. A combination of self-report and performance-based measures may maximize the information value as evidenced by prediction of hospital costs [42] and mortality [43].

**PRO Use in Clinical Trials as a Means to Gain Additional Psychometric Information**

Testing the PRO instrument in early clinical development phases of an intervention is valuable in terms of establishing its attributes in a setting most similar to the anticipated registration trial. The opportunity to include PRO measures within early phase trials may depend on whether the PRO is considered a primary or secondary end point. Phase I trials may not be optimal because study participants often do not suffer from the target condition, although exceptions exist, as in the case of cancer.

Phase II trials offer a better opportunity to assess how the instrument performs relative to a consistent dose and within a specific patient population.

Psychometric properties can be reevaluated and analysis methods explored to generate hypotheses for Phase III. The Phase III population may differ from the Phase II population. If broader, one must rely on the evidence from other reports to assess how this will affect the PRO assessment. If narrower, PRO data from a subgroup of the Phase II population may be reevaluated. Nevertheless, sample sizes in Phase II trials may range from 40 to 200 patients, and this small number may limit subgroup evaluations. If multiple Phase III trials are conducted, they can provide additional information about the psychometric properties of a PRO instrument.

**Evaluation of Psychometric Properties in Phase II Trials**

Phase II trials appear to be the most obvious choice for exploring the concept of interest and the associated validation issues. Phase II trials may allow for the evaluation of the PRO relative to other clinical measures. These can contribute to estimating the minimally important difference for the PRO instrument in a specific target population and interventional setting. In short, Phase II may provide the opportunity for hypothesis generation and for considering sample sizes for Phase III trial design.

**Evaluation of Psychometric Properties in Parallel with Phase III**

When PROs are secondary end points in clinical trials, the progression of clinical development does not always allow for full evaluation of an instrument before starting Phase III research. In such a situation, the sponsor may be willing to assume the risk of conducting evaluation studies in parallel with Phase III or within the Phase III trial itself. This type of situation may arise when the need for a modification is identified in Phase II, but evaluation studies can not be completed before initiating Phase III. Although not ideal, this situation is common in drug development.

**Evaluation of PRO Properties in Phase IV Studies**

Phase IV studies that further explore the PRO may be useful in further establishing PRO reliability and validity. Demonstrating adequate psychometric performance for any PRO supports the data as meaningful in the real world. Clinical trials are criticized as being artificial and not based on patient-experienced reality, because trials are conducted under strict conditions (tight enrollment standards, randomized, placebo-controlled, double-blinded) to minimize bias and minimize patient variability. These heavily orchestrated operations are costly and time consuming; yet they remain the standard by which regulatory decisions are made. PRO instruments that can withstand the rigors of such an operation gain credibility over time as tools in the decision-making process.
Eventually these tools can find their way into the more common clinical vernacular of Phase IV postmarketing trials. The therapeutic intervention is subject to wider ranges of bias and variability to reflect what happens in the real world. In these settings, effectiveness (defined as the balance of risk vs. therapeutic benefit) may be more important than efficacy (defined as the ability of the drug to show statistical superiority or noninferiority to a comparator). Therefore, a PRO might take on more of a lead role as the primary outcome domain.

An example of the use of a Phase IV study is found in the sleep literature [44]. In the context of excessive daytime sleepiness (EDS) in normal subjects versus EDS in narcoleptics, the investigator provided sensitivity- and specificity-based evidence that supported the abandonment of the historical “gold standards,” the Multiple Sleep Latency Test (MSLT) and the Maintenance of Wakefulness Test (MWT), in favor of the Epworth Sleepiness Scale (ESS; a simple eight-item, self-administered questionnaire). The investigator demonstrated that the MSLT and MWT were not as valid as the ESS because the latter better discriminated the EDS of narcolepsy from the EDS of normal subjects.

Summary

Table 1 summarizes the considerations for evaluating reliability and validity of PROs. Demonstrating reliability and validity is essential for determining whether a specific PRO measurement will be useful in the evaluation of a health-care intervention. These psychometric properties are on a continuum in which more evidence provides greater confidence in the value of the PRO data. The level of support for psychometric properties of a concept should be more extensive when the application presents greater stakes or potential consequences to individuals.

Proof of reliability is essential to ascertain that one is consistently measuring the same concept or dimensions of a single concept. For clinical trials, the 0.70 minimum reliability threshold is appropriate. To translate findings to the individual, higher reliabilities are preferred. Carrying equal importance is validity, which is ascertaining that the concept of interest is the one being measured. Construct validity will generally be an acceptable approach with PRO instruments. Content validity is important but only provides beginning evidence of validity. Criterion validity is difficult as rarely is there a PRO “gold standard” that can be used.

Several guidelines are recommended for establishing sufficient evidence for reliability and validity. To provide initial estimates of reliability and validity, samples sizes should be at least 200 cases and reliability and validity should be replicated in at least one additional representative sample. Subsamples may be created for cross-validation. Existing instruments, including item banks, should have at least one full report on their development and evaluation and one report on their use in an interventional setting. If the population characteristics or manner in which the PRO is administered vary from how the investigator intends to use the PRO, then additional psychometric information will be needed. To gain additional validity data, qualitative approaches such as focus groups, behavioral coding, and cognitive interviewing or the quantitative approach of DIF testing may be useful. If reliability and validity have not been adequately evaluated before Phase III testing begins, the sponsor assumes considerable risk in not being able to substantiate the use of the PRO. Phase II trials are an appropriate venue for evaluating psychometric properties for a target population and minimally important differences.

Source of financial support: Ron D. Hays was supported in part by the UCLA/DREW Project EXPORT, National Institutes of Health, National Center on Minority Health & Health Disparities (P20-MD00148-01), the UCLA Center for Health Improvement in Minority Elders/Resource Centers for Minority Aging Research, National Institutes of Health, National Institute of Aging (AG-02–004), and the National Institute of Aging (P01-AG-02–079). Funding for the Mayo/FDA meeting was provided by the Mayo Foundation in the form of unrestricted educational grants; North Central Cancer Treatment Group (NCCTG) (CA25224-27) and Mayo Comprehensive Cancer Center grants (CA15083-32).

References

1 ERIQA Group. Comments on the Reflection paper on the regulatory. Guidance for the Use of Health-

Table 1  Reliability and validity of PROs: summary points

| • Usefulness of a PRO measure depends on reliability and validity |
| • Reliability and validity lie on a continuum such that more evidence produces more confidence in PRO data |
| • Reliability indicates that the same concept or dimensions of a single concept is consistently measured |
| • Minimum reliability threshold for clinical trials is 0.70 |
| • Higher reliability of a PRO instrument will decrease sample size needed to detect a change in the PRO |
| • Validity indicates that the concept of interest is being measured |
| • Construct validity is a generally acceptable approach with PROs |
| • Ideally, sample sizes of at least 200 should be used to provide initial estimates of reliability and validity |
| • To gain additional validity data several qualitative and quantitative approaches are available |
| • Qualitative—focus groups, behavioral coding, cognitive interviewing |
| • Quantitative—item response theory, differential item functioning |
| • PRO item banks are a valuable resource for building standardized and linked short-form instruments or computer-adaptive tests. Their psychometric properties should be evaluated in ways similar to those used to evaluate any other PRO instrument. |
| • Phase II trials are useful for identifying small, but important differences |
| • Phase III need to be powered for PRO differences |
| • In Phase IV trials PROs often become more important than in earlier phase trials |

PRO, patient-reported outcome.


40 Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: true differences or differential item functioning. Medical Care 2003;41:III75–86.


