Improving Patient Reported Outcomes Using Item Response Theory and Computerized Adaptive Testing

ELIZA F. CHAKRAVARTY, JAKOB B. BJORNER, and JAMES F. FRIES

ABSTRACT. Objective. Patient reported outcomes (PRO) are considered central outcome measures for both clinical trials and observational studies in rheumatology. More sophisticated statistical models, including item response theory (IRT) and computerized adaptive testing (CAT), will enable critical evaluation and reconstruction of currently utilized PRO instruments to improve measurement precision while reducing item burden on the individual patient. Methods. We developed a domain hierarchy encompassing the latent trait of physical function/disability from the more general to most specific. Items collected from 165 English-language instruments were evaluated by a structured process including trained raters, modified Delphi expert consensus, and then patient evaluation. Each item in the refined data bank will undergo extensive analysis using IRT to evaluate response functions and measurement precision. CAT will allow for real-time questionnaires of potentially smaller numbers of questions tailored directly to each individual’s level of physical function. Results. Physical function/disability domain comprises 4 subdomains: upper extremity, trunk, lower extremity, and complex activities. Expert and patient review led to consensus favoring use of present-tense “capability” questions using a 4- or 5-item Likert response construct over past-tense “performance” items. Floor and ceiling effects, attribution of disability, and standardization of response categories were also addressed. Conclusion. By applying statistical techniques of IRT through use of CAT, existing PRO instruments may be improved to reduce questionnaire burden on the individual patients while increasing measurement precision that may ultimately lead to reduced sample size requirements for costly clinical trials.

Key Indexing Terms: OMERACT ITEM RESPONSE THEORY COMPUTERIZED ADAPTIVE TESTING PATIENT REPORTED OUTCOMES DISABILITY PHYSICAL FUNCTION

Measurement of patient reported outcomes (PRO) has been on the ascendancy in medicine, often led by rheumatologists, for nearly 25 years. These outcomes, recorded by patients on validated questionnaires such as the Health Assessment Questionnaire (HAQ)\(^1\), the Medical Outcomes Study 36-question Short Form (SF-36)\(^2\), and others, have become critical outcomes in both clinical trials and long-term observational studies in rheumatic diseases\(^3\)-\(^5\).

PRO include physical function or disability, side effects, medical care costs, pain, and other content areas. Instruments for measuring PRO are easier to administer and less expensive than physician-observed health status measures\(^6\). Well developed PRO instruments have been proven to be reliable, valid, and sensitive to change, and often have these attributes in greater abundance than do physician-reported measures. They exemplify outcomes according to the OMERACT filter of truth, discrimination, and feasibility\(^7\). Accordingly, they have become standard in evaluation and approval of new therapeutics, in observational studies, and in randomized trials. They have been translated and culturally adapted to scores of languages and cultures, and validated in thousands of studies\(^8\).

After a quarter-century, the time has come to critically evaluate the limitations of these standard measurements with the goal of improving the precision, ease of use, and responsiveness of these important PRO. One major limitation is the “one size fits all” approach by current standard tools. Since patients differ in symptoms and level of health, standard questionnaires contain many items that are irrelevant and uninformative for the particular patient. For any given patient, the clinician would probably prefer to ask more questions that are relevant for the patient’s level of health and to drop other questions. This may be achieved by developing a comprehensive bank of questionnaire items to measure the latent trait of interest (e.g., physical function/disability), use of a psychometric technique called item response theory (IRT)\(^9\) to place each item on a common ruler, and use of computerized adaptive testing (CAT) to administer the items. This way, with a similar or decreased questionnaire burden on participants, we will be able to more precisely measure the latent trait, and util-
mately, to reduce the sample sizes required for randomized clinical trials of new therapeutic agents. Physical function is the natural starting place for implementing these new techniques in musculoskeletal diseases, although the estimation of the latent trait may easily apply to other conditions.

An ambitious effort to institutionalize this approach is currently under way in the Patient Reported Outcomes Measurement Information System (PROMIS). This project is part of the US National Institutes of Health (NIH) “Roadmap” initiative and is designed to provide improved assessment of health status across all chronic illnesses as part of an improved infrastructure for clinical science. PROMIS is developing large patient-based item banks of hundreds of items with the aim of improving measurement instruments, reducing sample size, and bringing PRO to the level of the individual patient. We will describe the PROMIS approach to item bank development using physical function to illustrate the process.

The PROMIS approach to item bank development

The first step in the development of a comprehensive item bank is to construct a domain hierarchy that proceeds from the more global constructs (health) to more and more specific domains (physical, mental, and social health) to eventually lead to individual items (Figure 1). Each specific level of the hierarchy can be collapsed into the more general domain under which it falls. The hierarchy was developed by a modified Delphi technique involving more than 30 experts from 7 PROMIS centers. After initial general consensus of the 3 main domains was determined, an iterative process ensued to further develop more specific subgroups within each domain. The process of expanding the hierarchy map to the more and more specific constructs continues until the goal is reached of having the domains be mutually exclusive and collectively exhaustive, achieving a comprehensive unidimensional continuum that represents the latent trait of interest.

Alternative but compatible domain hierarchies have been developed by other groups, e.g., the model developed by the World Health Organization in its International Classification of Functioning, Disability, and Health (ICF) system. This comprehensive biopsychosocial framework of over 1400 categories distinguishes body functions, body structure, activities and participation, and personal and environmental factors.

The PROMIS domain hierarchy is a useful model for establishing the concepts for which item banks are developed. The arduous process of item bank development with PROMIS begins with collecting all individual proposed items from all known instruments in order to encompass the breadth of items in a manner that reduces bias. For the domain of physical function, 1860 items were collected from 165 English-language instruments. Each item in the large pool is reviewed in a structured qualitative process. The initial review is performed by at least 3 trained raters using defined criteria to eliminate items that are redundant, imprecise, sloppy, grammatically incorrect, directed at an inappropriately high reading/comprehension level, or contain inappropriate response options. This was followed by a modified Delphi approach to expert consensus, “Physical Function/Disability” was conceptualized as containing the 4 subdomains of “upper extremity,” “trunk,” “lower extremity,” and “complex activities” (instrumental activities of daily living; IADL). The subdomains identified through the iterative Delphi technique are encompassed within the Brief ICF Core Set for rheumatoid arthritis and were independently confirmed using more computational principle components analysis. Similar structures have now been confirmed by others.

Similar to the process of validating the ICF core set for rheumatoid arthritis, the reduced item pool for physical function/disability then undergoes extensive testing by patients through the use of focus groups, cognitive interviews, and patient surveys. Patients are asked to rate the importance of the item and clarity of the item, and are asked to describe the idiomatic meaning of the items.

Analyses of both expert and patient review of items have yielded several important concepts. Physical function as a latent trait is the ability to carry out activities of daily living (ADL) and instrumental activities of daily living (IADL). Items that address capability (“Are you able to...?”) more clearly address this latent trait than do performance questions (“In the last week, did you...?”) as the latter require opportunity as well as capability. This is illustrated by potential differences in answers to the major “strong grip” item: the question “Are you able to use a hammer to pound a nail?” may yield very different responses than “In the last week, have you used a hammer to pound a nail?” This variance in response is not due to differences in the underlying latent trait, and thus violates the assumption of unidimensionality. Other insights yielded from extensive expert and patient item review include the preference to use the present-tense timeframe for questions rather than referring to a specific timeframe; and use of 4- or 5-item “capability” responses to items including “normally,” “with some difficulty,” “with a little difficulty,” “with much difficulty,” or “unable to do,” while avoiding reference to time in response options (all of the time, some of the time, etc.) for physical functioning items. Last, attribution to disease or other limiting context has been eliminated from each item as it adds another, unwanted dimension: patients’ opinions regarding attribution as well as capability. Akin to the ICF hierarchy, we believe that these instruments should remain neutral to or independent of etiology of the level of the latent trait. The remaining item pool is then further refined to standardize response categories and item wording, and to fill any gaps or omissions with new items, particularly to address concerns about floor and ceiling effects.

Quantitative item evaluation and item response theory

Once the item pool has been reduced in number and revised through qualitative item review (to roughly 200 questions), the process of quantitative item review using item response
theory techniques begins\textsuperscript{17,18}. Item response theory (latent trait theory) has been used for over half a century in the fields of psychometrics and educational testing. The goals of IRT are to analyze a series of categorical variables (items) to precisely estimate a quantitative attribute, or latent trait (i.e., intelligence, depression, physical function).

Several key assumptions underlie IRT models: unidimensionality and local independence\textsuperscript{19} as well as the particular form of the IRT model itself. Unidimensionality refers to the concept that a group of items measures a single latent trait, and that all variance in responses is due to individual differences in the underlying latent trait. Local independence means that aside from the latent trait of interest, responses to any 2 items are statistically independent. The precision and robustness of a set of items to reliably estimate a given trait relies on the fulfillment of these assumptions and on the quality and comprehensiveness of the item bank.

The simplest IRT statistical model originates from the...
work of Rasch and focuses on the relative difficulty of items\textsuperscript{20,21}. The partial credit model, an extension of the Rasch model, is used when responses to a given item are on a multi-point scale, such as a Likert scale\textsuperscript{22}. In IRT models, the measurement properties of each item are evaluated by generating item response functions/characteristic curves that evaluate the probability of selecting any given answer to the item by the estimate of the latent trait. Based on the item response functions, an item information function and standard error of measurement (SEM) curves can be constructed for each item and used as an assessment of the precision of that item (Figure 2). In contrast to classical test theory, item information functions are not single coefficients, but rather functions that describe responses at different levels of the latent trait.

Item information functions can be summed to form scale information functions, which can then be transformed to SEM curves. In this manner, the SEM or measured precision of known (HAQ-DI and SF-36–physical function) and newly created instruments can be compared\textsuperscript{23}. These curves can be used to evaluate the precision of measurement of a given instrument over a wide range of estimates for the latent trait of interest. Different scales of the same construct can thus be compared and cross-calibrated at any given level of the latent trait of interest (Figure 3).

IRT approaches in rheumatology have usually tried to minimize the number of questions asked while maintaining or improving measurement characteristics of the instrument. The purpose of an improved disability assessment instrument, however, must be primarily to increase the precision of the estimates, at both the individual patient and the group level, while retaining face validity and the values of the patient\textsuperscript{24}. From increased precision comes increased sensitivity to change, leading to smaller and less expensive studies with the same statistical power. This process requires more questions, not fewer.

**Computerized adaptive testing**

CAT is a natural extension of IRT and is currently the most effective mechanism for achieving a high degree of measurement precision with a relatively brief questionnaire. It requires, in effect, that each patient be administered his or her own personal questionnaire, which in turn requires a computer or hand-held device\textsuperscript{18,24,25}. Within each content area, such as physical functioning, fatigue, or pain, CAT employs a form of artificial intelligence that selects questions with difficulty levels directly tailored to the test-taker, and shortens or lengthens the test to reach the desired level of precision. Good function leads to harder questions, poor function to easier ones, in an iterative process\textsuperscript{26}. The process continues until predefined stopping rules are satisfied and the chosen precision is

![Figure 2. IRT models evaluate the measurement properties of each item. Top panels show the probability of each of 3 possible responses based on the underlying latent trait/physical function score for 2 items in the SF-36 physical function questionnaire (PF01 and PF10). Bottom panels show the corresponding item information function curve derived from the response probabilities. The standard error of measurement curve is calculated as the reciprocal of the square root of the item information function.](image-url)

---

Chakravarty, et al: Improving patient reported outcomes

Copyright © 2007. All rights reserved.
obtained. In practice, this approach minimizes the number of items that need to be administered to an individual to obtain an estimate of functioning in any particular content area. It requires large item banks, not small ones, and patient-specific use of items. Because each item narrows the range of estimates of the latent trait, CAT assessments can achieve much higher precision than fixed questionnaires with the same numbers of items.27

The increasing computer and Internet accessibility and literacy of the general public and the increasing sophistication of hand-held computer devices has enhanced the feasibility of using IRT/CAT in outcome assessment both for clinical research and for use in daily clinical practice. Time and expense of mailing and scanning fixed questionnaires will be eliminated, and “point-and-click” features of Web or hand-held-based questionnaires will facilitate ease of use. CAT has the further advantage of being able to identify “aberrant” responses based upon the probability of any individual response to an item given the estimated physical function derived from prior questions. For example, if a person who is able to walk a mile responds that she is unable to walk a block, the computer will identify the response to this item as inconsistent with response patterns at the same level of physical function. These can be brought to the attention of the respondent for correction or clarification in real time. In this way, an additional level of quality control is built into the algorithm, further improving the accuracy of the estimates, and saving time previously required for manual quality control measures.

**DISCUSSION**

Patient-reported outcomes are among the main measures of clinical efficacy of therapeutic agents in clinical trials of rheumatic diseases. The statistical and technological sophistication is now available to take PRO to the next level of precision, while reducing the length of items posed to the individual respondent. The use of exhaustive, comprehensive item banking with IRT models applied through CAT will hopefully yield improved techniques for outcome assessment that may eventually supersede existing instruments and set new standards. The added ease of use and shortened individual questionnaire burden may enhance use of PRO in routine clinical care. Among the advantages these techniques possess is the ability to conduct detailed evaluations of each item, near elimination of floor and ceiling effects, and real-time quality control. Scores may be estimated from subsets of the items in the complete item bank, and different contemporary instruments may be cross-calibrated to achieve a common metric of the same latent trait. Newly developed items can be easily incorporated into the item bank with no need to change the entire metric. CAT offers further improvement by implementing user-friendly Web-based or hand-held modalities that enable rational selection of a short form of optimal items tailored to the individual. Inconsistent responses can be reevaluated or clarified in real time. Taken together, these techniques offer more precise and appropriate measures of patient reported outcomes, and with increased precision comes the potential to reduce sample size in clinical trials while retaining appropriate statistical power.
REFERENCES


