Qual Life Res (2011) 20:1159–1167
DOI 10.1007/s11136-011-9863-1

The king's foot of patient-reported outcomes: current practices and new developments for the measurement of change

Richard J. Swartz · Carolyn Schwartz · Ethan Basch · Li Cai · Diane L. Fairclough · Lori McLeod · Tito R. Mendoza · Bruce Rapkin · The SAMSI Psychometric Program Longitudinal Assessment of Patient-Reported Outcomes Working Group

Accepted: 21 January 2011 / Published online: 19 February 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract
Purpose Assessing change remains a challenge in patient-reported outcomes. In June 2009, a group of psychometricians, biostatisticians, and behavioral researchers from other disciplines convened as a Longitudinal Analysis of Patient-Reported Outcomes Working group as part of the Statistical and Applied Mathematical Sciences Institute Summer Psychometric program to discuss the complex issues that arise when conceptualizing and operationalizing "change" in patient-reported outcome (PRO) measures and related constructs. This white paper summarizes these issues and provides recommendations and possible paths for dealing with the complexities of measuring change.
Methods/Results This article presents and discusses issues associated with: (1) conceptualizing and operationalizing change in PRO measures; (2) modeling change using state-of-the-art statistical methods; (3) impediments to detecting true change; (4) new developments to deal with these challenges; and (5) important gaps that are fertile ground for future research.
Conclusions There was a consensus that important research still needs to be performed in order develop and refine high-quality PRO measures and statistical methods to analyze and model change in PRO constructs.

Keywords Outcome assessment (Health Care) · Quality of life · Longitudinal studies · Psychometrics · Statistical models · Response shift

Richard J. Swartz and Carolyn Schwartz contributed equally. Other authors listed alphabetically. Additional working group members include: Thomas Atkinson, Ph.D., Ken Bollen, Ph.D., Charles Cleeland, Ph.D., Cheryl Coon, Ph.D., Betsy Feldman, Ph.D., Theresa Gilligan, M.S., Herle McGowan, Ph.D., Knashawn Morales, Sc.D., Lauren Nelson, Ph.D., Mark Price, M.A., M.Ed. Bryce Reeve, Ph.D., Carmen Rivera-Medina, Ph.D., Quiling Shi, Ph.D., Rochelle Tractenberg, Ph.D., MPH, Xiaojing Wang, Jun Wang, and Valerie Williams, Ph.D.

R. J. Swartz (✉)
Jones Graduate School of Business, Rice University, Houston, TX, USA
e-mail: rswartz@rice.edu

C. Schwartz
DeltaQuest Foundation, Inc., Concord, MA, USA

C. Schwartz
Tufts University Medical School, Boston, MA, USA

E. Basch
Memorial Sloan-Kettering Cancer Center, New York, NY, USA

L. Cai
University of California, Los Angeles, CA, USA

D. L. Fairclough
University of Colorado Denver, Denver, CO, USA

L. McLeod
RTI Health Solutions, Research Triangle Park, NC, USA

T. R. Mendoza
The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA

B. Rapkin
Albert Einstein College of Medicine, Bronx, NY, USA

Introduction

Understanding the patient experience at multiple time points provides investigators and clinicians with a more comprehensive picture of the impact of disease and treatment over time. Substantial advancement has been made to create reliable, valid, and responsive instruments for measuring PROs. Determining how to best measure and quantify change in PROs over time is still under development. Historically, well-defined physical measures such as *length* had their developmental periods. What is now the standard "foot" in English Units was once a variable and uncertain measure. In ancient times, measures of length were based on body parts [1]; the traditional belief being that a person measured distance by the number of his own feet that would cover the length in question. Traditional accounts relay that later the length of the King's foot was declared the standard "foot", but could change with new kings. Finally, a unified, non-changing standard was adopted to create the modern definition.

In June 2009, a group of biostatisticians, psychometricians, and behavioral researchers from various other disciplines convened in a Longitudinal Analysis of Patient-Reported Outcomes Working Group as part of the Statistical and Applied Mathematical Sciences Institute Summer Psychometric program to discuss the complex issues that arise when conceptualizing and operationalizing "change" in PRO measures and related constructs. This white paper summarizes the issues discussed, reviews the current state of the art bringing together research from different disciplines such as psychometrics, statistics and psychology, and provides recommendations and possible paths for dealing with the complexities of measuring change. A more detailed report can be found online [2]. This white paper will discuss issues and recommendations associated with the following: (1) conceptualizing and operationalizing change in PRO measures; (2) modeling change using state-of-the-art statistical methods; (3) impediments to detecting true change; (4) new developments to deal with these challenges; and (5) important gaps that are fertile ground for future research (see Fig. 1).

Conceptualizing and operationalizing change in PRO measures

Classical test theory and item response theory

The paradigm that dominates the concept of change in psychometric research derives from *classical test theory* (*CTT*) [3]. In CTT, observed measures of change (based on subtraction of post-test from pretest scores on a given repeated measure) can be decomposed into change in an attribute's true score plus differences due to random error of measurement that occur at each observation [3–5]. Cronbach and colleagues [6] recognized that this error could distort true change and devised an approach to regress individuals' observed change scores toward the grand mean of sample change, in accord with the unreliability of the measure. Regression approaches have also been used to adjust for measurement-ceiling and measurement-floor effects that necessarily attenuate change (i.e., initial scores close to a measure's maximum value cannot increase as much as other scores). These classical

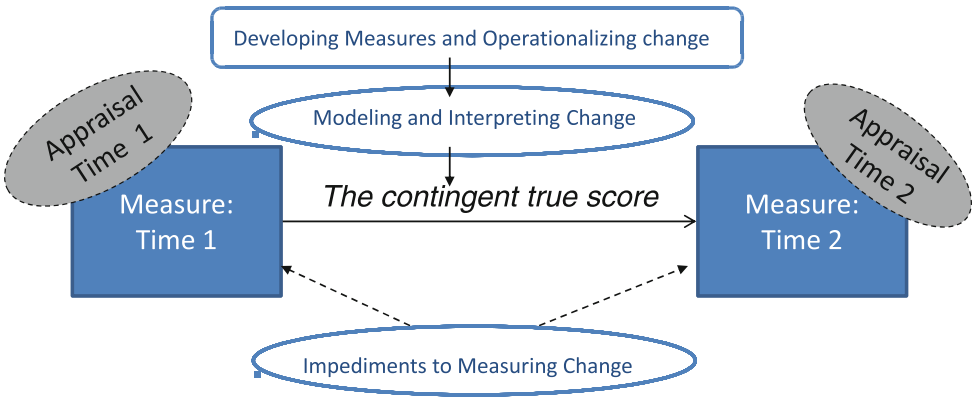

Fig. 1 Roadmap: This figure graphically describes the content of this white paper. Only two time points are considered for simplicity. Measurement occurs at 2 time points (time 1 and time 2). The *boxes* or *circles* that have no fill color represent major concepts/sections of the paper. *Dotted lines* indicate concepts that are rarely measured or accounted for. Section "Conceptualizing and operationalizing change in PRO measures" speaks to *developing measures.* Section "Modeling change using state-of-the-art statistical methods" of the paper discusses *modeling and interpreting change* using the developed measures. Section "Impediments to detecting true change" discusses *impediments to measuring change.* Specifically, this paper reviews *the contingent true score* model and how it can facilitate understanding change in PRO scores. The observed change depends on the measures at the two time points. Each measurement at each time point is influenced by an individual's *appraisal* parameters which may or may not be constant across the time points

corrections to change scores are problematic because they are highly dependent upon the characteristics of a given sample.

Recently, quality of life (QOL) research has benefited from seminal work done in educational testing using *item response theory* (IRT) methods [3, 7, 8]. In IRT, the estimation of an individual's *latent score* is based on a probabilistic model derived from the individual's responses to items with well-defined parameters of *difficulty* and *discrimination*. Using IRT, unbiased estimates of these parameters are attainable even when the sample may not be fully representative of the population. This makes them more consistent on a population level than their CTT counterparts [9].

To be included in a measure, item response probabilities must fit a specified model (usually a logistic function), as a function of the score on the latent trait of interest. Many of the mainstream IRT models do not allow for large effects of individual differences other than influences from the person's trait level. This approach is potentially problematic for PROs because relevant content may not be included if the items perform differently across group membership (e.g. gender) or because of unmeasured individual differences.

Operationalizing instrument responsiveness to true change

Measuring change requires instruments that are sensitive enough to detect that change. A PRO instrument is *responsive* if it shows change when there is true change (cf [10, 11]). Responsiveness is a contextualized attribute of an instrument rather than an unvarying characteristic—it is a function of who is being analyzed (individuals or groups), which scores are being contrasted (cross-sectional versus longitudinal), and what type of change is being quantified (observed change versus important change) [12]. A recent review of the available responsiveness statistics identified one index, Cohen's effect size [13], as most appropriate [10]. Cohen's effect size anchors observed change against variability at baseline, is less vulnerable to extreme values, and is more readily interpretable [10].

The 2009 FDA guidance for industry, *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims* [14] extends the responsiveness property to include quantifying a benchmark for change on the PRO scale which characterizes the meaningfulness of an individual's response (rather than a groups response) to active treatment. This requires determining the smallest PRO score difference that can be judged as meaningful. Various methods exist for estimating this minimal important difference (MID) for PRO scales (cf [15]). There are three common categories: (1) *Anchor-based methods* compare changes in PRO scores over time with patient- or clinician-reported global ratings of overall change in disease severity on a balanced Likert-type scale. For more detail, see the study by Juniper et al. [16] (2) *Distribution-based methods* rely on the distribution of the empirical data from an administration of the PRO measure. Most commonly the distribution-based MIDs are some function of the standard deviation of the baseline scores (this includes methods based on the Standard Error of Measurement). For more details, see Norman et al. [17] and Wyrwich et al. [18] (3) *Statistical Rules of Thumb* methods are, as the name implies, based on statistical rules of thumb. For example, across numerous studies a 0.5-point change or greater per response on a 7-point graded-response question has been applied to define an MID [19, 20]. More details regarding this method and the previous two are also given in the online report [2].

Patients achieving the benchmark response are considered treatment "responders," whereas patients not achieving that amount of response are considered "non-responders." The guidance provides examples of benchmarks, such as a 2-point change on an 8-point scale or a pre-specified percent change from baseline. These pre-specified values are defined using external measures and should be at least as large as an MID because minimal change may not be sufficient to classify a patient as a responder.

There is currently no consensus for defining an optimal MID value. Common practice is to calculate multiple estimates of MIDs and consider all their values to judge the meaningfulness of reported change [14, 21, 22]. After the range of MIDs has been defined, patient- or group-level comparisons may be made using the MID(s) as a guideline. For example, the percentage of patients achieving change of at least one MID for each domain can be compared across treatment groups as a criterion for the amount of improvement, and this can be statistically tested to determine if there is a more efficacious treatment.

Modeling change using state-of-the-art statistical methods

There are two general statistical model formulations that developed recently and somewhat independently that are now becoming widely used for analysis of longitudinal PRO data. The first is multi-level modeling (MLM). Such models are also called hierarchical linear models in educational and behavioral sciences [23, 24], and mixed-effects models or mixed models in biometrics and medical statistics [25]. Although we discuss linear models, there are also more complicated non-linear models [26]. The second modeling framework is structural equation modeling (SEM), particularly the latent curve models for repeated measures data [27]. SEM represents the culmination of econometric/sociometric

simultaneous equations (path) analysis and the psychometric factor analytic measurement models [28].

In MLM and SEM, a model is specified directly on the repeated observations for each individual. Both frameworks offer more flexibility and model change over time more accurately than methods based on Generalized Linear Models [29, 30]. MLM and SEM allow for individual differences in the initial status and rate of change, represented as (co)variance components. Both time-varying and time-invariant covariates can be included in the models to elucidate the causes and patterns of change for individuals and groups. Through the use of full-information maximum likelihood, MLM and SEM require only the relatively weak missing at random (MAR) assumption for missing data [31] (discussed in more detail in a section on missing data), and therefore, handle unbalanced designs. Importantly, MLM or SEM analyses provide estimates of individual characteristics of growth and how these individual characteristics relate to the covariates.

For a large class of models, MLM and SEM lead to equivalent model formulations. Essentially, the random effects—effects that are assumed to come from a distribution, as opposed to an unknown but estimable (fixed) constant—in MLM are specified as latent variables in SEM. MacCallum, Kim, Malarkey, and Kiecolt-Glaser [32] discuss these similarities in detail. If the two frameworks produce equivalent models, it can be shown that they lead to the same parameter estimates [33]. MLM is more advantageous when subjects are clustered (e.g., subjects nested within clinics) because modeling additional levels of nesting in MLM is straightforward. SEM is more flexible when some of the covariates are latent constructs that are measured by fallible observed indicators because SEM accounts for measurement error.

Impediments to detecting true change

Current methods allow one to identify what constitutes a true PRO score change and to model and interpret this change. However, there are threats to the measurement of true change. Three threats relevant to longitudinal PRO data are (1) response shift, (2) instruments with varying sensitivity across the trait of interest, and (3) non-ignorable missing data. Although all comparisons will be affected, estimates of intra-group change will be most strongly affected by these impediments.

Response shift

Individuals employ subjectivity to appraise their QOL and other PRO variables; in fact no measurement of QOL or related *evaluative constructs*—constructs whose measures depend on internal standards of the person reporting them and therefore have no external validations—is possible without subjective appraisal. An individual's criteria for these subjective constructs can change during a course of illness and treatment. Such "response shift" phenomena [34] are ubiquitous in health-outcome research, showing that individuals who experience health-state changes often modify their internal standards, values, and conceptualization of target constructs in an iterative process of adaptation [35].

These subjective aspects of a "response shift" add unmeasured variability into the model because many current PRO measurement instruments do not account for the subjective aspects that influence the score. These subjective aspects are potentially measureable and could be included as factors in the model. Perhaps a paradigm shift is required: instead of viewing response shift as an impediment to the measurement of change that must be prevented, the model is more appropriately considered misspecified. An important goal is to account for these appraisal factors when they pose a threat to measuring change.

To more fully address the nature of change in evaluative constructs, Schwartz and Rapkin introduced the notion of the *contingent true score.* They argue that any measure of an evaluative construct must be interpreted contingent on a cognitive-affective process of appraisal that underlies an individual's response [36]. They formulate that the true score depends upon four parameters of appraisal: *Frame of Reference,* the individual's frame of reference or interpretation of an item; *Sampling of Experience,* an individual's process of selecting experiences within the frame of reference; *Standards of Comparison,* the standards used to evaluate the experiences; and *Individuals Combinatory Algorithm,* the algorithm for combining or reconciling discrepant experiences and evaluations [36]. For evaluative measures, Rapkin and Schwartz hypothesize relationships between changes in appraisal parameters and the three types of response shift—reconceptualization, reprioritization, and recalibration [36]. Ethnographic methods by Wyrwich and colleagues [37] have shown that these four appraisal parameters substantially account for individuals' introspective statements about their QOL and help validate this relationship between appraisal and response shift.

Sensitivity changes across the PRO continuum

It is challenging to develop items to measure across the entire range of a trait, and PRO instruments tend to be adequately sensitive for only subsets of the range of the construct [38]. Even the newly developed Patient-Reported Outcomes Measurement Information System (PROMIS) instruments, whose item banks underwent extensive qualitative and quantitative development, tend to be more sensitive and reliable at certain values for the construct [39–41].

For example, the PROMIS Pain Impact bank has very little reliability at the lower end of the range (see http://www.assessmentcenter.net/ac1/). These qualitative issues affect the measurement of PROs, and affect the ability to measure PROs longitudinally. (see [38] for a more detailed discussion). Although not a complete solution, rigorous IRT analysis can facilitate assessing the varying sensitivity of the instrument to identify instruments most sensitive to the change of interest in a particular study.

Missing data

Missing data causes challenges when assessing change in PRO scores. Often the data are missing because of the unobserved outcome; for example, the subject's QOL has dramatically declined, and the subject stopped reporting their data. The missingness of the data depends on the unobserved outcome, and such missingness cannot be ignored. Such data are non-ignorable missing data, or missing not at random (MNAR). These data are problematic because selection of the appropriate analytic models depends on untestable assumptions [42–44]. (cf [43] for helpful review). As most statistical methods assume that missing data are missing at random (MAR)—that the probability of missingness depends only on the observed data—the typical approach for non-ignorable missing data is to collect or determine ancillary data that capture the relationship between missingness and the unobserved data. For a helpful review of classical (multiple imputation) and modern (pattern mixture) methods for handling missing data, see [43, 44]. For more details see the online report [2].

The problems of non-ignorable missing data and response shift have some similarities; both are situations in which ancillary data improve the likelihood that the estimation of change is closer to the true change. The challenge in both areas is to identify the ancillary variables through research and to incorporate them prospectively into clinical investigations.

New developments

New developments in PRO methodology address some of the previously mentioned challenges. New modeling techniques enhance the PRO measures and address other issues related to change, such as identifying an MID. Assessing appraisal variables and accounting for their effects have been shown to increase the sensitivity to detecting true change [45].

Multidimensional IRT

The advantages of IRT over CTT in health outcomes measurement have been demonstrated by a recent wave of applications [46]. Mainstream IRT assumes the underlying construct that the items measure can be represented with a single (unidimensional) latent trait. It is becoming more evident that several PRO measures and specifically general QOL measures are more accurately modeled as a combination of constructs. This suggests a multidimensional model might fit better. If response shift occurs, or if appraisal information is to be included in QOL or PRO measures, a multidimensional model might better capture both the PRO construct and appraisal characteristics. Multidimensional IRT relaxes the unidimensional assumption so that the observed items are influenced by multiple latent variables [47]. In longitudinal settings, multidimensional IRT models can facilitate detection of response shift or inclusion of appropriate appraisal information in PRO measures.

Consider a PRO construct (i.e. depression) that is being assessed pre- and post-treatment for a group of respondents. The relationship between the items and the latent construct may change over time. From a psychometric perspective, this is a classical longitudinal measurement invariance issue. Modeling and testing invariance requires an IRT equivalent of the longitudinal factor analysis model [48]. In this model, the unidimensional construct at the two time points is represented as two correlated factors, each measured by the time-specific item responses. Also additional latent variables can be included to model residual dependence across time induced by using the same item twice.

Figure 2 shows the general structure of this model for three items and two occasions. Rectangles represent the observed items and circles represent the latent variables.

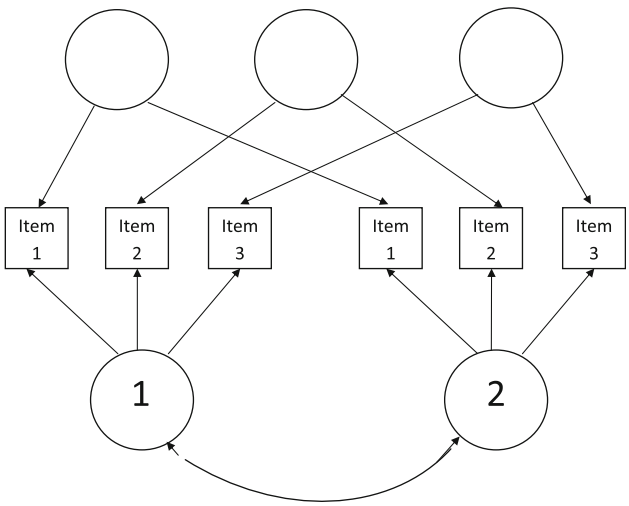

Fig. 2 IRT equivalent of the longitudinal factor analysis model using 3 items and 2 measurement occasions. *Rectangles* represent observed items (items 1–3); *circles* represent latent variables. *1* represents the latent variable at time 1, and *2* represents the latent variable at time 2. These are represented as 2 separate correlated factors. The *blank circles* represent additional latent factors that model residual dependence that can occur when the same item is used repeatedly over time

Such a longitudinal multidimensional IRT model was considered by Hill for dichotomous responses, and by te Marvelde, Glas, & van Damme, with some restrictions, for ordinal responses [49, 50]. By manipulating the constraints on the item parameters in a multidimensional IRT model, researchers can tease apart the observed change into two components: change due to response shift and true change in the level and variability of the latent construct [51].

Widespread use of multidimensional IRT models is hampered by challenging computational issues in parameter estimation. Recent computational advances in adaptive quadrature [52], Markov chain Monte Carlo methods [53], and stochastic approximation [54, 55] are poised to resolve them. Cai [51] developed a two-tier item factor analysis modeling framework and efficient computational tools that facilitate longitudinal IRT analysis and tests of longitudinal measurement invariance. Further developments such as this are required to make multidimensional IRT more accessible to applied researchers.

New methods for MID

New methods have been developed to define MID values by incorporating clinical or patient-based judgment. The "consensus" value approach requires that a panel of healthcare professionals or patient panelist determine an MID based on a series of appraisals using their clinical experience and standard-setting techniques [56]. Another approach combines clinical judgment and receiver operating characteristic (ROC) curves to identify the unit of change on the PRO that best predicts clinical judgment of the minimal important change [57]. Finally, one can use IRT to compute change in the PRO measure on the standardized latent construct scale, building upon the anchor-based methods by identifying the minimal value in terms of the minimally identified change on the latent construct [58].

The QOL appraisal profile

Assessment of changes of appraisal can account for response shift and provide a more complete understanding of PRO scores. Rapkin and Schwartz introduced a QOL Appraisal Profile that can elicit multiple measures of each of the four appraisal parameters [36]. Li and Rapkin [59] have demonstrated that this tool can be useful for examining response shift in QOL in HIV patients. A rich area of future research involves how to incorporate appraisal information into PRO measurement.

Consensus, recommendations, and future directions

Identifying a consistent and replicable "king's foot" measure of true PRO change is challenging because the process of self-report is complex and there are many subjective factors that can influence self-report scores [36, 60, 61]. Much progress has been made in measuring and analyzing PROs longitudinally; nevertheless, new research understanding the role of appraisal and response shift, as well as how to operationally define MIDs, is still required.

Most PRO measures do not account for the appraisal factors that affect how people report their experience. Developing measures to incorporate appraisal parameters can greatly reduce the obfuscating effect of a response shift and increase measurement precision. It is as yet unclear how to introduce appraisal assessment to the measurement of change in PROs.

Appraisal parameters necessarily influence how individuals respond to and evaluative items. Item evaluation and selection according to IRT assumptions must then account for and reflect appraisal processes. It is necessary to investigate what frame of reference we are unintentionally imposing by excluding conceptually relevant items with unacceptable fit to IRT models or high differential item functioning across diverse groups. In addition, current approaches to assess appraisal involve qualitative data that require considerable coding for quantitative analysis. Indeed, Bloem suggests that appraisal parameters are best assessed for each individual item [62]. It would be desirable to identify a parsimonious system to describe and quantify appraisal parameters for use in further quantitative analysis.

Finally, in the scenario described in the *Multidimensional IRT* section, response shift is still seen as a type of nuisance, and not as an intrinsic part of measurement for PROs. To capture true change in PROs, we recommend that future research bring together concepts from IRT and the contingent true score model. First it is necessary to understand when it is important to measure appraisal and when it is not. The more evaluative a measure is, the more potential importance appraisal can play. This will require better measurement of appraisal phenomenon, and expanding current mainstream IRT models not only to include multiple latent variables (such as current multidimensional IRT models) but also including additional explanatory information into the IRT models. Rijmen, Tuerlinks, De Boeck, and Kuppens [63] presented IRT models as a special case of non-linear mixed models and De Boeck and Wilson [4] further expanded that relationship into a framework they call Explanatory Item Response Models. The benefits of such a framework is the ability to handle measurement issues such as the multiple latent traits (multidimensionality) as well as statistical or predictive issues such as developing models that adjust for differential item functioning, modeling changes in appraisal parameters, and understanding the effect of appraisal parameters on measurement.

It is also worth mentioning that technological advances such as computerized adaptive testing and interactive voice response methods offer promising areas of future research to overcome many of the issues with longitudinal PRO measurement such as missing data issues, and possibly facilitate standardization of some of the appraisal variables. More discussion can be found in [2]. Using IRT (or multidimensional IRT) and related computer-adaptive testing approaches to assess appraisal itself may be useful to better assess appraisal parameters while minimizing the burden associated with such assessment. Critical future research investigating explanatory item response models for use in computerized adaptive testing would further facilitate such approaches to incorporating appraisal. Because appraisal processes are subject to change over time, measuring "true change" in evaluative measures is difficult without at least a thorough understanding of the appraisal process and its effect on measurement.

Another approach might be to adapt PRO measures to include new items or new collection techniques that can anchor some of the important appraisal variables. For example, during a chemotherapy treatment regimen, one can fix the appraisal variables at specific levels such that they remain consistent for patients across the treatment period. This requires understanding the appraisal process for instrument development, but would not necessarily require measuring appraisal with the hypothetical PRO instrument.

Despite notable developments in our understanding of MID and responsiveness, additional work in interpreting change values is required. Current practice uses patient-level units to make both patient-level and group-level judgments. Smaller group-level changes or smaller differences in means between groups may equate to similar treatment impacts as indicated by patient-level classifications based on an MID. Comparing patient groups who "achieved MID" versus those who "did not achieve MID" by treatment is not equivalent to testing whether the *means* of the treatment groups are statistically different.

Second, some MID approaches (described in more detail in [2]) use only a portion of the available data. Data from a sometimes very small subset are used to estimate the MID, while the remaining data are ignored or simply checked to ensure that the pattern of change matches the expected change on the PRO scale. Future research should further develop methods to incorporate all the data from those methods to define MIDs or ways of interpreting change.

Third, MIDs are applied uniformly along the entire score range. Under certain circumstances it may be more reasonable that MIDs vary according to disease severity or some other characteristic—minimal important changes may depend on where a patient *begins* on the concept measured (similar to an ANCOVA-type argument). In addition, the measurement precision of the instrument should also be taken into account when defining an MID. For example, an MID should not be identified and then applied at the patient level if it is less than the standard error of measurement.

Finally, very little work has been done related to the contingent true score model and its effect on developing MID estimates. Integrating the contingent true score model and MID forces one to revisit the definition and interpretation of MID and to reconsider response shift effects on the MID. Rich areas of future research in the longitudinal analysis of PROs involve continued investigation of appraisal variables, new methods to identify responders, methods to determine responsiveness, methods to improve estimates in the presence of non-ignorable missing data and even reduce missing data.

Acknowledgments The authors would like to thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) for supporting the working group. The authors would also like to acknowledge editorial critique from Lee Ann Chastain and William Carmichael. The first author would like to note that he recently changed positions and can now be contacted at Rice University, but most of the work in this manuscript occurred while he was at The University of Texas M. D. Anderson Cancer Center and his funding came from Career development grant K07-CA-113641 (Principal Investigator Richard J. Swartz, Ph.D.) from the National Cancer Institute. Li Cai's research is made possible by grants from the Institute of Education Sciences (R305B080016, Principal Investigator Li Cai, Ph.D., and R305D 100039, Principal Investigator Noreen Webb, Ph.D.) and grants from the National Institute on Drug Abuse (R01DA026943 Principal Investigator Maria Orlando Edelen, Ph.D. and R01DA030466 Principal Investigator Li Cai, Ph.D.). Tito R. Mendoza and Charles S. Cleeland were funded in part by grants from the National Cancer Institute: RO1CA026582 (Principal Investigator Charles S. Cleeland Ph.D.) and 5P01CA124787 (Principal Investigator Charles S. Cleeland, Ph.D.). The authors also thank the two reviewers and the Editor for their helpful critiques and comments.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Brief History of Measurement Systems with a Chart of the Modernized Metric System. (1974). In N. B. O. Standards (Ed.), Washington, DC: Government Printing Office.
2. Swartz, R. J., Basch, E., Cai, L., Fairclough, D. L., McLeod, L. D., Mendoza, T. R., Rapkin, B., Schwartz, C. E., & The SAMSI Psychometric Program Longitudinal Assessment of Patient-Reported Outcomes Working Group. (2010). The king's foot of patient-reported outcomes: Current practices and new developments for the measurement of change. from http://www.bepress.com/mdandersonbiostat/paper60.
3. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Pub. Co.
4. de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

5. McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.
6. Cronbach, L. I., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74*(1), 68–80.
7. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
8. van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
9. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.
10. Norman, G. R., Wyrwich, K. W., & Patrick, D. L. (2007). The mathematical relationship among different forms of responsiveness coefficients. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16*(5), 815–822.
11. Fayers, P. M., & Hays, R. D. (2005). *Assessing quality of life in clinical trials: Methods and practice* (2nd ed.). Oxford: Oxford University Press.
12. Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemilology, 54*(12), 1204–1217.
13. Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.
14. US Department of Health and Human Services. (2006). *Food and drug administration*. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims draft guidance.
15. de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes, 4*, 54.
16. Juniper, E. F., Guyatt, G. H., Willan, A., & Griffith, L. E. (1994). Determining a minimal important change in a disease-specific quality of life questionnaire. *Journal of Clinical Epidemiology, 47*(1), 81–87.
17. Norman, G. R., Sloan, J., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality-of-life. *Medical Care, 4*, 582–592.
18. Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of Clinical Epidemiology, 52*(9), 861–873.
19. Guyatt, G. H., Juniper, E. F., Walter, S. D., Griffith, L. E., & Goldstein, R. S. (1998). Interpreting treatment effects in randomised trials. *British Medical Journal, 316*, 690–693.
20. Norman, G. R., Sridhar, F. G., Guyatt, G. H., & Walter, S. D. (2001). Relation of distribution- and anchor-based approaches in interpretation of changes in health related quality of life. *Medical Care, 29*, 1039–1047.
21. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109.
22. Sloan, J., Cella, D., & Hays, R. D. (2005). Clinical significance of patient-reported questionnaire data: Another step towards consensus. *Journal of Clinical Epidemiology, 58*, 1217–1219.
23. Raudenbush, S. W., & Byrk, A. S. (2002). *Heirarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
24. Reise, S. P., & Duan, N. (2003). *Multilevel modeling: Methodological advances, issues, and applications*. Mahwah, N.J.; London: Lawrence Erlbaum Associates.
25. Demidenko, E. (2004). *Mixed models: Theory and applications*. Hoboken, NJ: Wiley.
26. Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.
27. Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. Hoboken, NJ: Wiley.
28. Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
29. Hedeker, D. R., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.
30. Muller, K. E., & Stewart, P. W. (2006). *Linear model theory: Univariate, multivariate, and mixed models*. Hoboken, NJ: Wiley.
31. Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
32. MacCallum, R. C., Kim, C., Marlarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32*, 215–253.
33. Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*, 135–167.
34. Ahmed, S., & Schwartz, C. E. (2009, in press). Quality of life assessments and response shift. In A. Steptoe (Ed.), *Handbook of behavioral medicine: Methods and applications*. New York, NY: Springer Science.
35. Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science & Medicine (1982), 48*(11), 1507–1515.
36. Rapkin, B. D., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health and Quality of Life Outcomes, 2*, 14.
37. Wyrwich, K. W., & Tardino, V. M. (2006). Understanding global transition assessments. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 15*(6), 995–1004.
38. Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 25–46.
39. Garcia, S. F., Cella, D., Clauser, S. B., Flynn, K. E., Lad, T., Lai, J. S., et al. (2007). Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative.[erratum appears in J Clin Oncol. 2008 Feb 20;26(6):1018 Note: Lad, Thomas [added]]. *Journal of Clinical Oncology, 25*(32), 5106–5112.
40. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology, 61*(1), 17–33.
41. Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16*(Suppl 1), 109–119.
42. Fairclough, D. L., Thijs, H., Huang, I. C., Finnern, H. W., & Wu, A. W. (2008). Handling missing quality of life data in HIV clinical trials: What is practical? *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 17*(1), 61–73.
43. Hogan, J. W., Roy, J., & Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine, 23*(9), 1455–1497.
44. Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., & Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine, 21*(8), 1023–1041.
45. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587–598.

46. Edelen, M. O., Reeve, B. B., Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(Suppl 1), 5–18.
47. Edwards, M. C., & Edelen, M. O. (2009). Special topics in item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (pp. 178–198). New York, NY: Sage.
48. Tisak, J., & Meredith, W. (1989). Exploratory longitudinal factor analysis in multiple populations. *Psychometrika, 54*, 261–281.
49. Hill, C. D. (2006). *Two models for longitudinal item response data.* Unpublished doctoral dissertation, University of North Carolina—Chapel Hill.
50. te Marvelde, J., Glas, C. A. W., Landeghem, G., & van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement, 66*(1), 5–34.
51. Cai, L. (2010). A Two-Tier Full-Information Item Factor Analysis Model with Applications. *Psychometrika, 75*(4), 581–612.
52. Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533–555.
53. Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541–561.
54. Cai, L. (2010). High-dimensional Exploratory Item Factor Analysis by a Metropolis-Hastings Robbins-Monro Algorithm. *Psychometrika, 75*(1), 33–57.
55. Cai, L. (2010). Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis. *Journal of Educational and Behavioral Statistics, 35*, 307–335.
56. Harding, G., Leidy, N. K., Meddis, D., Kleinman, L., Wagner, S., & O'Brien, C. (2009). Interpreting clinical trial results of patient-perceived onset of effect in asthma: Methods and results of a Delphi panel. *Current Medical Research and Opinion, 25*(6), 1563–1571.
57. Turner, D., Schünemann, H. J., Griffith, L. E., Beaton, D. E., Griffiths, A. M., Critch, J. N., et al. (2009). Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *Journal of Clinical Epidemiology, 62*(4), 374–379.
58. McLeod, L. D., Nelson, L., & Lewis, C. (July 1, 2009). Personal Communication.
59. Li, Y., & Rapkin, B. (2009). Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of Clinical Epidemiology, 62*(11), 1138–1147.
60. Schwartz, C. E., & Rapkin, B. D. (2004). Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health and Quality of Life Outcomes, 2*, 16.
61. Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (2000). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Lawrence Erlbaum.
62. Bloem, E. F., van Zuuren, F. J., Koeneman, M. A., Rapkin, B. D., Visser, M. R. M., Koning, C. C. E., et al. (2008). Clarifying quality of life assessment: Do theoretical models capture the underlying cognitive processes? *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 17*(8), 1093–1102.
63. Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185–205.