Published in final edited form as:
Expert Rev Pharmacoecon Outcomes Res. 2011 December ; 11(6): 677–684. doi:10.1586/erp.11.74.

Advancing PROMIS's methodology: results of the Third Patient-Reported Outcomes Measurement Information System (PROMIS®) Psychometric Summit

Adam C Carle[*], **David Cella**,
Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Li Cai,
Department of Education and Psychology, University of California Los Angeles, Los Angeles, CA 90095-1521, USA

Seung W Choi,
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 710 N. Lake Shore Drive, Chicago, IL 60611, USA

Paul K Crane,
Internal Medicine, School of Medicine and Health Services, School of Public Health, University of Washington, Box 359780, 325 Ninth Avenue, Seattle, WA 98104, USA

S McKay Curtis,
Department of Statistics, University of Washington, Seattle, WA 98104, USA

Jonathan Gruhl,
Department of Statistics, University of Washington, Seattle, WA 98104, USA

Jin-Shei Lai,
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 710 N. Lake Shore Drive, Chicago, IL 60611, USA

Shubhabrata Mukherjee,
General Internal Medicine, University of Washington, Box 359780, 401 Broadway, Suite 5076, V126, Seattle, WA 98104, USA

Steven P Reise,
Chair, Quantitative Psychology, University of California Los Angeles, 3587 Franz Hall, Los Angeles, CA 90095, USA

Jeanne A Teresi,
Columbia University Stroud Center and New York State Psychiatric Institute, Research Division, Hebrew Home at Riverdale, 5901 Palisade Avenue, Riverdale, NY 10471, USA

David Thissen,

[*]Author for correspondence: University of Cincinnati School of Medicine and University of Cincinnati College of Arts and Sciences, James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7014, Cincinnati, OH 45229, USA, Tel.: +1 513 803 1650, Fax: +1 513 636 0171, adam.carle.cchmc@gmail.com.

Financial & competing interests disclosure

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Department of Psychology, University of North Carolina Chapel Hill, CB#3270, Davie Hall, Chapel Hill, NC 27599-3270, USA

Eric J Wu, and
Psychology Department, University of California Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095, USA

Ron D Hays
Department of Medicine, University of California Los Angeles, 911 Broxton Avenue, Los Angeles, CA 90095-1736, USA

Abstract

In 2002, the NIH launched the 'Roadmap for Medical Research'. The Patient-Reported Outcomes Measurement Information System (PROMIS®) is one of the Roadmap's key aspects. To create the next generation of patient-reported outcome measures, PROMIS utilizes item response theory (IRT) and computerized adaptive testing. In 2009, the NIH funded the second wave of PROMIS studies (PROMIS II). PROMIS II studies continue PROMIS's agenda, but also include new features, including longitudinal analyses and more sociodemographically diverse samples. PROMIS II also includes increased emphasis on pediatric populations and evaluation of PROMIS item banks for clinical research and population science. These aspects bring new psychometric challenges. To address this, investigators associated with PROMIS gathered at the Third Psychometric Summit in September 2010 to identify, describe and discuss pressing psychometric issues and new developments in the field, as well as make analytic recommendations for PROMIS. The summit addressed five general themes: linking, differential item functioning, dimensionality, IRT models for longitudinal applications and new IRT software. In this article, we review the discussions and presentations that occurred at the Third PROMIS Psychometric Summit.

Keywords

computerized adaptive testing; dimensionality; factor analysis; item response theory; patient-reported outcomes; PROMIS; psychometrics; structural equation modeling

---

To improve clinical research and health outcomes in the USA, the NIH developed its 'Roadmap for Medical Research'. The Roadmap identified high-priority scientific opportunities and needs that the NIH should pursue and that no institute could accomplish singly [1]. The Patient-Reported Outcomes Measurement Information System (PROMIS®) is one of the Roadmap's key developments. To create the next generation of patient-reported outcome (PRO) measures, PROMIS utilizes item response theory (IRT) [2] and computerized adaptive testing (CAT) [3].

IRT uses probabilistic, mathematically based models to describe how people tend to respond to questions. It models the relationship between individuals' responses to questions about their health and the underlying (i.e., latent) variable measured by an instrument. As a result, it offers several advantages over classical test theory [4]. For example, IRT allows more accurate and realistic estimates of reliability by allowing precision to vary across scores. It can create shorter yet more reliable instruments (particularly when coupled with CAT). In addition, it lets one compare different scales on a linked, common metric. All of these advantages directly support PROMIS's goals, which include developing the next generation of PROs that (with minimum questions) provide meaningful, precise measurement, while simultaneously requiring fewer respondents to achieve statistical power in research settings.

To continue advancing PROMIS's goals and achievements (described in detail at [101]), NIH funded PROMIS's second wave of studies (PROMIS II) in 2009. PROMIS II studies continue to advance PROMIS's general agenda, but also include new features (e.g., longitudinal data collection and more sociodemographically diverse samples). Additionally, PROMIS II includes an increased emphasis on pediatric populations and on evaluation of PROMIS item banks for clinical research and population science. These aspects bring new psychometric challenges. To address this, investigators associated with each of the PROMIS sites gathered in September 2010 at the Third PROMIS Psychometric Summit to describe and discuss pressing psychometric issues and new developments in the field, as well as make analytic recommendations for PROMIS generally. The 2010 summit aimed to address 5 general themes: linking, differential item functioning (DIF), dimensionality, IRT models for longitudinal applications and new IRT software. In this article, we describe each of the 2010 Summit's presentations, as well as the summary recommendations resulting from the Summit.

Linking scores across scales

The focus on developing pediatric measures has resulted in increased attention to ensuring comparable scores across pediatric and adult scale forms. Thus, the Summit started with a discussion of linking. Generally, linking refers to statistical efforts to transform scores from one measure into the metric of another [5]. David Thissen noted that recent advances allow investigators to improve upon the two previously viable options (prediction and alignment) when linking measures developed separately and to different specifications, as with PROMIS pediatric and adult forms.

Investigators can now use a full-information factor analytic approach to linking, called calibrated projection [6,7]. Calibrated projection eases the requirement of previous full-information methods that the two tests measure a single construct. In calibrated projection, one fits a multidimensional IRT (MIRT) model to the two measures' item responses. The model describes two distinct but correlated variables as representing the underlying constructs measured by each scale. The approach also allows for additional latent variables (e.g., nuisance variables) or correlated errors to address local dependence(s) that may exist. Thissen recommended calibrated projection over previous methods that required that each scale measure a single construct given its flexibility and the fact that it uses all of the information in the item response matrix [8], as opposed to limited-information estimation based on joint frequencies [9]. This provides an advantage because full-information estimation is less sensitive to chance occurrence of large inter-item correlations, which can excessively influence limited-information estimation. Thissen's discussion ended with a description of the planned linkage of the PROMIS pediatric and adult scales. Adolescents and young adults will complete eight PROMIS pediatric short forms that have adult item bank analogues. This project will evaluate whether the linkage appears useful (i.e., does linking provide reliable score estimates?). If so, the project will then employ either unidimensional IRT calibration and calibrated projection as appropriate to align the pediatric and adult scales.

To make concrete PROMIS's linking goals (and hurdles), David Cella and colleagues described their efforts to calibrate new item banks that extend existing item banks. Jin-Shei Lai described the preliminary results of a study seeking to link PROMIS measures with related scales in order to expand the range of options for PROs. By linking new to existing item banks, the project will allow an expanded range (e.g., larger item pool) of PRO assessment options on a standardized metric. Lai's examples included adult (n = 1316: 803 general population, 513 cancer patients) and pediatric (n = 513) examples; both used Stocking–Lord [10] separate calibrations and fixed parameter calibration and included

sufficient participants to meet sample size requirements [11]. Preliminary findings indicated that in both examples, separate and fixed methods tended to produce similar parameter results and linking did not provide new information across levels of the latent trait where information was previously scarce. However, linking did yield a somewhat more peaked information function in previously well-measured trait levels.

Given these findings, some questioned the utility of expanding existing item banks if the results did not substantially increase the ability to measure different construct level. In response, several participants noted that an assessment of a linking procedure depends partly on the intended use of the item bank. Linking improved the amount of information in the existing range, leading to more accurate estimates. Importantly, while the examples may not have shown substantial improvements or differences, that need not be the case. Expanding the existing PROMIS item banks by linking additional items could lead to notable improvements for other constructs by including more content and coverage of the measured constructs.

Differential item functioning

PROMIS seeks to generate items that produce equivalent measurement across diverse sociodemographic groups (e.g., race/ethnicity). DIF, sometimes referred to as measurement bias and item bias, refers to the possibility that two individuals with equivalent health tend to respond to questions about their health differently as a function of another variable. For example, DIF would exist if (on average) two people with the same level of physical functioning rated their physical functioning differently as a function of their age. DIF limits the ability to make reliable and valid cross-group comparisons. Introducing the DIF presentations, Jeanne Teresi emphasized the importance of examining DIF [12] and noted that increased sociodemographic diversity in the PROMIS II samples will increase DIF analyses opportunities. She then presented general DIF analyses guidelines [13]. These include: qualitative analyses and cognitive interviews; generating DIF hypotheses; examining items and raw scales to detect distributional skew and sparse data (combining categories as necessary); selecting anchor items in advance if possible (excluding items with DIF when necessary); examining model fit and assumptions; performing DIF analyses and sensitivity analyses; examining DIF's impact at aggregate and scale levels; and performing DIF adjustments if possible and feasible.

The Summit included detailed presentations of three DIF methods. Using a common dataset, each presentation focused on DIF across age on the short form of the PROMIS Physical functioning items in a sample of older adults. The data included three assessments of physical functioning at baseline (n = 521), 6 months and 1 year [14].

Analyzing the baseline data, Adam Carle presented multiple-indicator multiple-cause (MIMIC) models [15,16]. MIMIC models describe a measurement model (in the factor analytic tradition) to describe how people tend to respond to questions. MIMIC models include a covariate (i.e., source of DIF) to examine the direct and indirect effects of the covariate on measurement. Direct effects describe DIF as traditionally considered: the covariate directly leads to differential likelihood of item endorsement at the same trait level. Indirect effects refer to the possibility that, because the covariate may influence the measured variable itself, it may indirectly influence responses. Results found DIF across several items. For four items, DIF analyses indicated that, at the same level of the latent trait, older individuals tended to report worse physical functioning than younger individuals. For one item asking about the ability to wash and bathe one's self, analyses indicated that older individuals (relative to younger) tended to endorse better physical functioning at the same level of the latent trait.

Emphasizing the need to evaluate the extent to which identified DIF influences substantive conclusions, Carle showed the results of models ignoring and incorporating DIF. Prior to accounting for DIF, older adults endorsed significantly poorer physical functioning. Yet, after adjusting for item-level DIF, the two groups did not differ significantly, suggesting that DIF had a meaningful impact. Surprisingly, this adjustment would result in a conclusion opposite expectations (i.e., that older individuals do not differ from younger individuals in physical functioning). To provide a simple example of MIMIC, Carle noted that he had not included multidimensionality in the MIMIC model, which could have led to the surprising conclusion. He and others used this to emphasize the importance of developing the most accurate model possible in order to make appropriate DIF conclusions.

Carle noted several of MIMIC's strengths. MIMIC does not require categorizing a continuous variable in order to examine DIF across groups. Additionally, because MIMIC analyses work with a pooled covariance matrix, they can examine DIF in smaller sample sizes. However, MIMIC also has weaknesses. Chiefly, one typically cannot examine DIF in the loadings/discriminations (nonuniform DIF). Other methods overcome this limitation.

A general latent variable approach offers greater flexibility in DIF analyses than MIMIC. Rich Jones expanded the MIMIC model to a multiple group (MG)-MIMIC model that described separate physical functioning measurement models for each age group and included time points as covariates. This model examined DIF across age and time. These analyses found nonuniform DIF in five items, only two of which were in common with the MIMIC results. Unlike the MIMIC impact analyses, MG-MIMIC did not indicate that DIF caused erroneous conclusions. Thus, Jones highlighted the importance of interpreting DIF within the framework of the chosen and other potential models.

Jones identified the flexibility in models as one of the general latent variable modeling approaches' strengths. Strengths also included the ability to incorporate multiple variables simultaneously and handle longitudinal data. However, flexibility also serves as a weakness. One must carefully consider the data, likely causes of DIF, and develop specific *a priori* hypotheses. Importantly, the interpretation of DIF (as with any approach) depends upon the model specified. One must acknowledge that other models and methods might result in substantively different (and perhaps more valid) conclusions.

To offer another method, Mukherjee and Crane presented a longitudinal analysis based upon a hybrid ordinal logistic regression (OLR)/IRT framework [17]. This approach compares a hierarchically nested series of three OLR models that predict an item's response. OLR Model 1 predicts an item's response probability using IRT latent trait estimates, a term describing group membership, and a term describing the group by trait estimate interaction. OLR Model 2 estimates item response probability using only the latent trait estimates and the group term. OLR Model 3 estimates the probability solely from the IRT trait estimate. One then compares the differences in log likelihoods from Models 1 and 2 to a χ^2 distribution to identify nonuniform DIF (DIF in the discrimination parameters), and from Models 2 and 3 to identify uniform DIF (DIF in the location parameters). One does this for each of the set's items to probe for DIF in the item set.

Using the same data as the other DIF presenters, Mukherjee and Crane first performed cross-sectional analyses at each time point. At each time point, items identified with DIF differed. Subsequently, the authors extended the cross-sectional OLR/IRT approach to the longitudinal setting by accounting for covariance within individuals by clustering on person. When doing this, the authors found both uniform and nonuniform DIF, which manifested across different items across time. Like the MG-MIMIC analyses, the results suggested that,

while statistically significant DIF existed, it did not appear to substantively influence conclusions, suggesting the DIF was trivial.

The longitudinal OLR/IRT approach has several important features. It can account for the average effect of DIF across time, essentially adjusting for DIF at each time point to provide an unbiased estimate of change across time. It can also examine and describe changes in how DIF manifests across time. In addition, the approach does this making use of all the available data (allowing missing data across time) and using widely available software (e.g., PARSCALE).

Considerable discussion occurred regarding the choice of DIF method, given that each resulted in different conclusions. The presenters and participants noted that the choice of a 'best' DIF detection method depended on several factors, including the type of data, question of interest, reason for DIF and goal of detection. As with linking, the group concluded that one single approach would not prevail. Instead, a convergence of results using different approaches should be sought. One could view different methods as sensitivity analyses. Analysts will need to use and compare and contrast the results of several approaches, weighing the implications of each in light of the results and the methodological differences across the methods. Importantly, analysts should evaluate the practical impact of DIF.

Dimensionality

In psychometrics, dimensionality refers to the number of constructs data appear to measure. Multidimensionality (measuring multiple constructs) can profoundly affect IRT model parameters. Steve Reise noted that, although many IRT models assume unidimensional data, insufficient understanding exists regarding the assumption's implications, how to best assess unidimensionality and the implications of unidimensionality violations on IRT model parameters. Reise's discussion focused on the tension that exists when data fail to fit a truly unidimensional model. Both broad and narrow constructs exist, responses may measure a set of separable and important constructs, but even multidimensional data may yield scores influenced primarily by a single common factor [18].

Reise highlighted the importance of correctly accounting for violations of unidimensionality in order to correctly estimate IRT parameters. To date, analysts have traditionally evaluated the data to determine whether it demonstrated 'sufficient unidimensionality' and, if so, applied unidimensional IRT models. Analysts have used various indices of dimensionality to support this. However, forcing multidimensional data into unidimensional models will generally result in a mis-specified latent variable and spurious IRT parameters. As Reise illustrated, none of the fit indices used to 'justify' the unidimensional approach actually evaluate the degree to which an inappropriate unidimensional model results in distorted parameters. When faced with multidimensionality, investigators should compare the difference in the IRT parameters that result across unidimensional and bifactor models. Bifactor models generally describe models that specify a common general factor that influences all item responses, while also explicitly modeling 'grouping' or 'nuisance' factors [19] that influence small sets of item responses. By estimating exploratory bifactor models and comparing results to unidimensional IRT results, investigators can empirically evaluate multidimensionality's influence. At the very least, researchers should report multidimensional solutions so that readers can assess the tenability of a unidimensional model's results.

Discussion turned to handling multidimensionality in cross-sectional and longitudinal data and CAT. Most CAT applications assume and employ unidimensional IRT models [3] and measure individual traits one at a time. To achieve unidimensionality, researchers often

remove questions that lead to multidimensionality, leading to content under-representation. However, even after achieving unidimensionality, investigators often administer a battery of measures. Current CAT methods ignore multidimensionality and the information that would result from modeling multidimensionality. Seung Choi argued that a multidimensional and/or hierarchical CAT would provide a flexible modeling framework, content breadth, and capitalize on CAT's ability to reduce respondent burden.

Choi described a general multidimensional CAT (M-CAT) framework based on Cai's [7] two-tier full-information factor analysis approach, as implemented in IRTPRO (see below) [20]. M-CAT allows for multidimensionality and hierarchical structures, encompassing correlated traits, bifactor and two-tier hierarchical models. Initial results based on simulated and empirical data suggest: M-CAT performs at least as efficiently as unidimensional CAT but with enhanced validity (e.g., addresses conditional dependency issues); greater gains in efficiency occur with more highly correlated dimensions; and M-CAT requires content balancing to represent the dimensions more systematically and consistently using *a priori* target proportions. Choi showed that M-CAT captures nuisance dimensions relatively poorly and that it achieves maximal efficiency when focusing on hierarchical dimensions. While some issues remain unresolved, researchers wishing to measure several traits simultaneously or include hierarchical measures should consider M-CAT. However, the ability to employ M-CAT requires the ability to confidently identify the multidimensional structure.

McKay Curtis encouraged IRT analysts to adopt a Bayesian-averaging approach when attempting to model nuisance/grouping factors, residual covariation between items and secondary domains. Many tests, despite intending to measure a single construct, result in data that depart from unidimensionality and/or include local dependence. As described by Reise *et al.* [19], bifactor models can provide a parsimonious way to model these features. However, researchers rarely have clarity when selecting a bifactor model's secondary structure. Often investigators fit a variety of models, choosing the model that simultaneously provides the best fit and theoretical justification. Unfortunately, selecting and using a single model ignores uncertainty regarding the secondary factor structure.

Curtis proposed a Bayesian approach to account for secondary structure uncertainty. Bayesian approaches use probability distributions to model uncertainty. For example, in a bifactor analysis, an investigator would assign a prior probability distribution to model parameters. The prior distribution reflects uncertainty about the values before observing the data. Subsequently, the posterior distribution reflects updated knowledge about parameter values after observing the data. Bayesian approaches make inferences using the posterior distribution.

Curtis' method incorporated uncertainty about the secondary structure using a process [21] that generates a probability distribution (of structures) on a space of probability distributions. Given their discrete nature, random draws from the process will likely have repeated values [21]. In Curtis' approach, the repeated values form the secondary structure's basis, leading to his description of the model as a "random bifactor model". In his example, which used WinBUGS software [22] and data (n = 819) measuring cognitive functioning, Curtis showed that, while the random bifactor model assigned fairly large posterior probabilities to the model selected using the traditional approach, it also assigned relatively large probabilities to other feasible secondary structures. By including these probabilities in the posterior distribution, a researcher would arrive at more precise inference regarding the resulting item parameters and their standard errors.

IRT models for longitudinal applications

Meeting discussion included presentation of new IRT models for longitudinal data. Jonathan Gruhl described hierarchical Bayesian approaches to longitudinal item response models. As noted earlier, Bayesian methods acknowledge uncertainty about parameter values by specifying a prior distribution for parameters and updating the prior distribution by combing expectations with what the data reveal via a posterior distribution. As an additional and important advantage, Gruhl described how Bayesian models extend to hierarchical formulations, particularly beneficial to longitudinal data [23]. In these models, data are nested at several levels, which allows analysis at multiple levels and for the acknowledgement of uncertainty at each level.

In an example, Gruhl used longitudinal data and considered item responses nested within individuals nested within assessment occasions. The model considered item parameters invariant over time, but allowed the latent variable's mean to vary across time as a function of time-varying covariates and individual-specific effects. This formulation led to a time-varying mean structure for the latent trait and dependence among estimates of the latent trait for an individual over time. By extension, one could allow the mean function for the individual-specific coefficients to vary as a function of individual specific covariates and population parameters. One could also introduce individual-specific and item-specific random effects, which would induce additional within-individual within-item correlations; autoregressive terms in the mean function for the latent trait; and, finally, nonparametric terms. Bayesian analysts can use the widely available and free software program WinBUGS [22] (or any other general Markov Chain Monte Carlo software program). However, while Bayesian approaches offer potential, challenges remain in their implementation (e.g., how to specify the prior distribution; appropriate summaries of the posterior distribution; and computationally intensive Markov Chain Monte Carlo methods).

In a related vein, Li Cai described the application of a two-tier full-information factor analysis approach to longitudinal data. The two-tier item factor analysis approach represents a confirmatory (i.e., restricted) factor model. Essentially, the two-tier model imposes specific restrictions on the factor pattern and latent variable distributions that result in substantial computational advantages [7] but still allow sufficient flexibility to estimate numerous models. Importantly, the general two-tier model includes MIRT [24], bifactor models [25] and testlet models [26], and it generalizes bifactor and testlet models to include dichotomous, ordinal and nominal items. Because it allows two or more correlated primary factors in bifactor models, one would not have to break a larger scale into two separate bifactor models and estimate biased IRT parameters.

Cai noted that even in a simple longitudinal context, with the same scale used to measure a single unidimensional trait at two time points, the data's longitudinal nature creates a multidimensional structure. The model reflects a unidimensional structure within a given structure, but it requires at least two dimensions (one for each time point) to simultaneously model the data's longitudinal aspect and examine change in the means and variances of the constructs across time, as well as correlation across time. The two-tier model can capture these and model the conditional dependence that results from measuring the same individuals longitudinally. Moreover, the two-tier model allows other advances over more traditional approaches beyond those described in the longitudinal example (see [7] for details). Both the two-tier full-information factor analysis and Bayesian approaches will substantially open the field to new and exciting opportunities.

New IRT software

IRT software advances enable an improved ability to fit and test IRT models. The Summit included a detailed presentation of two: EQSIRT and IRTPRO. EQSIRT will address several important needs. First, structural equation modeling and IRT overlap. However, structural equation modeling typically uses means and covariances in estimation, while IRT traditionally uses full-information, response pattern-based estimation. Each approach has strengths and weaknesses. However, few IRT programs offer both approaches within a single program. This will increase the ease with which analysts can apply either. Moreover, this should expand the scope of IRT models to more readily include longitudinal and multilevel IRT models, as well as models with covariates. Second, EQSIRT will allow analysts to easily fit MIRT models, filling an important need in the IRT software field. Third, most currently available IRT programs (e.g., PARSCALE, BILOG and MULTILOG) are not user-friendly, nor do they incorporate many recent methodological advances in the field. EQSIRT will be packaged along with EQS and deliver a user-friendly interface. It will offer users a variety of methods in a convenient manner and provide graphical and tabular output easily integrated in research reports.

IRTPRO [20] implements the two-tier full-information factor analysis approach [7]. In addition, it also implements adaptive quadrature estimation [27] and the Metropolis–Hastings Robbins–Monro [28,29] algorithm to handle truly high- dimensional IRT models. IRTPRO will easily handle missing data (cross-sectionally and longitudinally), model multidimensionality and bifactor models, and address longitudinal questions. The program computes several model evaluation indices and has features to conduct DIF testing. The program allows user-defined restrictions on item parameters. This enables researchers to conduct likelihood ratio DIF tests, as well as Wald DIF tests. The likelihood ratio test mimics the IRTLRDIF anchoring method [30], which results in an item-by-item assessment of DIF. The Wald DIF test uses an ‘anchor-all’ ‘test-all’ method developed by Langer [102] in addition to anchored DIF. IRTPRO will compute EAP or MAP scores and standard errors. Like EQSIRT, IRTPRO should be available in 2012.

Conclusion

The PROMIS Psychometric Summit covered several key issues and hurdles that PROMIS investigators and IRT analysts will generally face in the coming years. Although not intended as a ‘cookbook’ for conducting psychometric analyses, several summary recommendations resulted from the Summit. First, analysts should always evaluate the extent to which data meet the assumptions of the selected IRT model, particularly in terms of the data’s dimensionality. When evidence for multidimensionality exists, investigators should recognize that ‘unidimensional enough’ does not address whether the failure to incorporate multidimensionality into the model would influence estimated item parameters. Researchers must evaluate the extent to which forcing multidimensional data into a unidimensional model causes meaningful distortion in the estimated item parameters. Second, while the two-tier and bifactor IRT methods do not solve all multidimensionality problems, analysts should take advantage of the computationally efficient shortcuts provided by the two-tier/bifactor models when the dimensionality structure is close to hierarchical. Relatedly, Bayesian random bifactor models, once fully vetted, offer analysts an excellent method for developing bifactor models when good *a priori* reasons do not exist for choosing a particular bifactor model. Bayesian models for longitudinal data offer similar strengths.

Third, with respect to linking, analysts should carefully consider the relation between scores on any pair of scales planned for linking to decide whether linked scores would be sufficiently precise to be useful for the planned purpose. Additionally, when linking,

analysts should always use an appropriate model (e.g., unidimensional or multidimensional). Fourth, when conducting DIF analyses (whether across age or other variables), investigators must acknowledge that multiple approaches exist. Given that the strengths and weaknesses of each differ, analysts should use more than one method and compare and contrast the results of each, interpreting heterogeneity (if any) in the results in light of the methodological differences across the methods.

In summary, the PROMIS cooperative group structure has enabled us to focus on advancing the measurement of health-related quality of life. A new generation of IRT analysis and software options is now available to study linking, DIF, dimensionality and longitudinal applications. While unresolved challenges remain (e.g., evaluating DIF with bifactor models), PROMIS investigators will continue to address challenging methodological and analytical issues as they present themselves along the way and, importantly, by doing so, the advances resulting from the PROMIS initiative will improve the quality, integration and application of PROs in the medical field.

Five-year view

Within 5 years, well-developed, psychometrically sound PROs resulting from the PROMIS initiative will have continued to become key outcome variables in clinical trials and practice. This will result through the application of IRT and the methods described in our paper. Within 5 years, investigators will routinely and more robustly evaluate IRT assumptions (e.g., dimensionality). They will regularly establish the validity of measures across heterogeneous groups. They will consistently utilize advances in the field that now solve many (though not all) dimensionality problems when they exist. They will employ Bayesian techniques that will have become more fully integrated into IRT and, as a result, they will more routinely use CAT to precisely and concisely measure PROs.

Key issues

- Numerous analytical challenges exist when seeking to precisely and concisely measure patient-reported outcomes across the demographically heterogeneous pediatric and adult health populations.
- Among them, analysts must better evaluate and model the extent to which data do or do not violate fundamental item response theory (IRT) assumptions.
- Investigators should use two-tier, bifactor and/or Bayesian IRT methods to solve assumption violations that can arise.
- When attempting to compare scores from disparate instruments, researchers should carefully consider whether linked scores would provide sufficient precision for this purpose.
- Additionally, they should always use an appropriate IRT model when linking scores across instruments.
- Investigators should establish the validity of instruments across heterogeneous groups using multiple approaches.
- The Patient-Reported Outcomes Measurement Information System initiative collaboratively uses and creates advances in IRT to meet these challenges.

Acknowledgments

Adam Carle would like to thank Tara J Carle, Lyla SB Carle and Margaret Carle, whose unending support and thoughtful comments make his work possible. Jeanne Teresi would like to thank Elayne Livote, Joseph P Eimicke, Marjorie Kleinman and Katja Ocepek-Welikson for processing and analyses of the data used in several presentations. Ron D Hays would like to thank Pam Hays for her support and encouragement during the trials and tribulations of PROMIS. He would also like to thank Karen L Spritzer for her computer programming support.

The following grants supported this work: NINR-R15NR10631 and 3U01AR057940-02S1 (AC Carle); 1U54AR057951 (D Cella); NCI-U01-AR057971, NIMHD, P60, MD00206, NIA, P30 and AG28741 (J Teresi); R305B080016, R305D100039, NIDA-R01DA026943 and NIDA-R01DA030466 (L Cai); NIA-R01 AG 029672 (PK Crane); National Institutes of Health through the NIH Roadmap for Medical Research Grant AR052177 PROMIS Project (R Hays) and 1U01AR052181-01 and 2U01AR052181-06 (D Thissen); and SBIR contract HHSN-2612007-00013C, National Cancer Institute (D Thissen).

References

Papers of special note have been highlighted as:

• of interest

•• of considerable interest

1. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Medical Care. 2007; 45 Suppl. 1(5):S3. [PubMed: 17443116] •• Provides an account of the Patient-Reported Outcomes Measurement Information System (PROMIS) I initiative and its accomplishments.
2. Hambleton, RK. Item Response Theory. USA: Kluwer, MA; 1985.
3. Van Der Linden, W.; Glas, C. Computerized Adaptive Testing: Theory and Practice. The Netherlands: Springer Netherlands; 2000.
4. Reeve BB. Item response theory modeling in health outcomes measurement. Expert Rev. Pharmacoeconomics Outcomes Res. 2003; 3(2):131–145. •• Offers an in-depth discussion of item response theory and its advantages.
5. Holland, P. Linking and Aligning Scores and Scales. Statistics for Social and Behavioral Sciences Part I. USA: Springer, NY; 2007. A framework and history for score linking; p. 5-30.
6. Thissen D, Varni J, Stucky B, Liu Y, Irwin D, Dewalt D. Using the PedsQL™ 3.0 athma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). Qual. Life Res. 2011; 20(9):1497–1505. [PubMed: 21384264]
7. Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika. 2010; 75(75):581–612. • Describes a key methodological advance that allows analysts to overcome commonly encountered problems when estimating item response theory models in real data.
8. Bock R, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika. 1981; 46(4):443–459.
9. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984; 49(1):115–132.
10. Stocking ML, Lord FM. Developing a common metric in item response theory. Appl. Psychol. Meas. 1983; 7(2):201.
11. Kolen M, Brennan R. Test Equating, Scaling, and Linking: Methods and Practices. 2004Berlin, GermanySpringer Verlag • Provides a detailed account of the issues that confront investigators when they seek to link different instruments to a common metric.
12. Teresi JA, Stewart AL, Morales LS, Stahl SM. Measurement in a multi-ethnic society. Overview to the special issue. Med. Care. 2006; 44 Suppl. 3(11):S3–S4. [PubMed: 17060831]
13. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. Psychol. Sci. Q. 2009; 51(2):148. [PubMed: 20336180]

14. Fries JF, Krishnan E, Rose M, Lingala BB, Bruce B. Improved responsiveness of physical function (disability) scales based upon item response theory. Arthritis Rheum. 2009; 60(Suppl.):S229.
15. Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. J. Am. Stat. Assoc. 1975; 70(351):631–639.
16. Carle A. Mitigating systematic measurement error in comparative effectiveness research in heterogeneous populations. Med. Care. 2010; 48(6):S68. [PubMed: 20473212] •• Discusses the importance of establishing that measures provide equivalent measurement across heterogeneous populations and discusses the application of an advanced model (multiple-group multiple-indicator multiple-cause) to address this issue.
17. Crane PK, Gibbons LE, Jolley L, Van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Med. Care. 2006; 44 Suppl. 3(11):S115–S123. [PubMed: 17060818]
18. Reise S, Moore T, Haviland M. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. J. Pers. Assess. 2010; 92(6):544–559. [PubMed: 20954056]
19. Reise S, Morizot J, Hays R. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual. Life Res. 2007; 16:19–31. [PubMed: 17479357] •• Explains the importance and application of the bifactor model.
20. Cai, L.; Du Toit, SHC.; Thissen, D. IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling. USA: Scientific Software International, IL;
21. Ferguson T. A Bayesian analysis of some nonparametric problems. Ann. Stat. 1973; 1(2):209–230.
22. Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 2000; 10(4):325–337.
23. Fox J, Glas C. Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika. 2001; 66(2):271–288.
24. Reckase, M. Multidimensional Item Response Theory. Berlin, Germany: Springer Verlag; 2009.
25. Gibbons R, Hedeker D. Full-information item bi-factor analysis. Psychometrika. 1992; 57(3):423–436.
26. Wainer, H.; Bradlow, E.; Wang, X. Testlet Response Theory and its Applications. Cambridge, UK: Cambridge University Press; 2007.
27. Schilling S, Bock RD. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. Psychometrika. 2005; 70(3):533–555.
28. Cai L. High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. Psychometrika. 2010; 75(1):33–57.
29. Cai L. Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. J. Educ. Behav. Stat. 2010; 35(3):307.
30. Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW., editor. Differential Item Functioning. USA: Lawrence Erlbaum Associates, NJ; 1993. p. 67-113.

Websites

101. PROMIS®. www.nihpromis.org.
102. Langer MM. A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation. 2008 http://dc.lib.unc.edu/cdm4/item_viewer.php?CISOROOT=/etd&CISOPTR=2084.