PROMIS®
Instrument Development and Validation
Scientific Standards
Version 2.0
(revised May 2013)

The Patient-Reported Outcome Measurement Information System (PROMIS®), funded by the National Institutes of Health, aims to provide clinicians and researchers access to efficient, precise, valid, and responsive adult- and child-reported measures of health and well-being. PROMIS instruments are based on modern measurement theory and include the rigorous application of quantitative, qualitative and mixed methods approaches for instrument development.

This document describes a set of standards that serve as the scientific foundation for the development and validation of PROMIS item banks and instruments. A general summary of instrument development and validation standards is followed by several appendices that outline specific components, followed by a final appendix that summarizes a maturity model for PROMIS item banks and instrument development and validation. The standards outlined in this document were based on the experience of PROMIS investigators, the published literature, and several existing sets of scientific methodology documents. These scientific standards are operationalized by a series of guidelines that provide detailed guidance for item bank development and validation, as well as summary of existing PROMIS practices. Reference citations are provided at the end of each individual guidance document. Similarly, the instrument maturity description document provides information concerning the readiness of measures derived from the PROMIS item banks for use in clinical research and practice.

Instrument development and validation is a process of accumulating evidence and, therefore, some standards, such as those related to translation, validity, reliability, or interpretability, will pertain only to those item banks and instruments that have a relevant language translation or have achieved levels of validity, reliability, or responsiveness. Similarly, not every PROMIS product is a calibrated item bank, hence guideline regarding item banks and CAT instruments have limited applicability to these sets of items.

List of Standards

1. Definition of Target Concept and Conceptual Model
2. Composition of Individual Items
3. Item Pool Construction
4. Determination of Item Bank Properties
5. Testing and Instrument Formats
6. Validity
7. Reliability
8. Interpretability
9. Language Translation and Cultural Adaptation

Scientific Standards

1. Definition of Target Concept and Conceptual Model

The conceptual model and target concept underlying the proposed instrument(s) should be defined and based on extant literature with input from content and measurement experts, clinicians, end-users, individuals (e.g. patients) and other respondents, as well as stakeholders as appropriate. In addition, the placement of the instrument within the PROMIS framework should be clearly defined.

Checklist:

1. Evidence that extant literature clearly informs model provided
2. Review by content and measurement experts conducted using sound qualitative approaches
3. Review by clinicians, patients and/or end-users conducted using sound qualitative approaches
4. Instruments placement in PROMIS domain framework specified

Related Guideline Documents:
Domain Framework and Definition (Appendix 2)
Qualitative Methods (Appendix 3)

2. Composition of Individual Items

Individual items should be refined through cognitive interviewing to ensure understanding and readability of item concepts, and be reviewed for translatability and literacy. In addition, consideration at the item level of both life course, and cultural harmonization should be addressed. Existing PROMIS item formats (stems, responses, tense and person) should be considered and utilized as appropriate. Otherwise, rationale for new item formats should be provided.

Checklist:

1. Used existing PROMIS item formats or provided rationale for item formats
2. Cognitive interviews conducted
3. Translatability review conducted
4. Cultural harmonization addressed
5. Literacy assessment conducted

Related Guideline Documents:
Structure & Composition of Individual Items (Appendix 4)
Qualitative Methods (Appendix 3)
Translatability & Cultural Harmonization Review (Appendix 5)
Literacy Review (Appendix 6)

3. Item Pool Construction

The item pool should provide measurement across a pre-specified breadth of the target construct. Adequate coverage of relevant sub-concepts (subdomains) that are considered important to the performance of fixed length forms and CATs derived from the item bank should be provided by the item pool. Construction of the item pool should also address issues

related to conceptual reconciliation with original proposed target construct and domain framework as needed. Cultural harmonization and life-course review conducted at the item pool level should be completed to ensure adequate coverage of cultural and development issues of the target construct. If possible, the item pool should cover a full breadth of the target construct as demonstrated by the list of the facets that were covered in development.

Checklist:

1. Rationale for inclusion or deletion of subsets of items or facets from conceptual perspective provided
2. Review of final item pool coverage of original target construct completed with reconciliation of the final and original target construct and PROMIS domain framework completed as needed
3. Cultural harmonization and life-course review conducted at the item pool level to ensure adequate coverage of cultural and development issues of the target construct

Related Guideline Documents:
Qualitative Methods (Appendix 3)
Intellectual Property (Appendix 7)

4. Determination of Item Bank Properties

The psychometric characteristics of the items contained within an item bank should be determined based on a representative sample of respondents and be demonstrated to have adequate measurement characteristics including dimensionality, model fit, and item and scale properties. Differential item functioning (DIF) for identified key groups (such as gender, age, education, race/ethnicity, language translation, literacy levels, diagnostic group) should be assessed (see maturation model), and the impact on measurement properties identified. If the set of items is not intended to be used as a calibrated item bank, a rationale, intended use, and appropriate measurement characteristics should be defined.

Checklist:

1. Dimensionality of the items within the item bank evaluated using appropriate statistical methods
2. Adequate item response theory model fit, including statistical assumptions necessary for IRT, demonstrated for the items within an item bank
3. Adequate item performance characteristics and scale performance characteristics demonstrated for the items within the item bank or set of items.
4. Differential item functioning (DIF) in key groups (age, gender, diagnostic grouping, education) assessed and the impact of DIF on measurement properties identified

Related Guideline Documents:
Measurement Model (Appendix 8)
Multi-dimensional IRT (Appendix 9)
Differential Item Functioning –Identification of DIF (Appendix 10)
Differential Item Functioning – Purification (Appendix 11)

5. Testing and Instrument Formats

Instrument formats should be appropriately defined based on intended use and item bank properties. Instrument formats may include CATs, fixed length short-forms, screener or profile formats. Instruments should demonstrate adequate scale properties and performance and include assessment of respondent burden. Instruments based on different modes (e.g. self-report, proxy-report, interview) and methods (e.g. computer, paper-pencil, telephone) of administration should have demonstration of comparable scale properties and performance and assessment of respondent burden for each mode.

Checklist:

1. Demonstration of adequate scale/ test-level properties of the instrument
2. Precision and efficiency of instruments identified across the measurement scale
3. Instrument performance parameters specified
4. Respondent burden characterized (in terms of time, number of items etc.)
5. Comparability of modes/methods of administration addressed

6. Validity

Construct, content and criterion validity should be addressed relative to a priori hypothesized relationships with related measures such as clinical indicators of severity or existing validated instruments of the target concept. The description of the methods and sample used to evaluate validity, including hypotheses tested and rationale for the choice of criterion measures, should be provided. The final instrument should be re-reviewed by experts and end-users/individuals to assess consistency with or identify differences between original definitions and final product.

If an instrument is purported to be responsive and/or intended to be used longitudinally, evidence or demonstration of adequate responsiveness based on relevant anchor-based methods in representative populations should be provided. Longitudinal data should be collected that compares a group that is expected to change with a group that is expected to remain stable. Rationale should be provided for the external anchors used to document change and the time intervals used for assessment.

Checklist:

1. Evidence supporting construct validity provided
2. Evidence supporting criterion validity provided
3. Evidence supporting content validity provided
4. Evidence supporting responsiveness provided

Related Guideline Documents:
Validity (Appendix 12)

7. Reliability

The reliability of the instrument should be described, including the methods used to collect data and estimate reliability. Internal consistency reliability estimates may consist of information and standard errors at different locations of the scale (item response theory) or reliability estimates and standard errors for all score elements (classical test theory). The reproducibility, or test-retest reliability, of the measure should be described, providing rationale to support the design of the study and the interval between first and subsequent administration to support the assumption that the population is stable.

Checklist:

1. Evidence supporting reliability across the target construct range provided
2. Evidence supporting test-retest reliability provided

Related Guideline Documents:
Reliability (Appendix 13)

8. Interpretability

The interpretation of instrument scores should be described, that is, the degree to which one can assign easily understood meaning to the instrument's quantitative scores. Rationale should be provided for the external anchors used to facilitate interpretability of scores. Information should be provided regarding the ways in which data from the instrument should be reported and displayed. The availability of comparative data from the general population and/or age-, gender-, or other group-specific scores should be described. Guidance should be provided regarding meaningfulness of scores and changes in scores for use by researchers and clinicians (e.g., minimally important differences, responder analyses).

Checklist:

1. Evidence supporting interpretation guidelines (MID and responder criteria) provided
2. Normative, comparative or reference data provided

9. Translation and Cultural Adaptation

If translated into another language, translation of items and instruments should include both forward and backward translations of all items and response choices as well as instructions. Translation of items, response choices and instructions should be obtained through an iterative process of forward and back-translation, bilingual expert review, and pre-testing with cognitive debriefing. Harmonization across all languages and a universal approach to translation should guide the process.

Checklist:

1. Items, response choices and instructions translated using a rigorous translation process

Related Guideline Documents:
Translation and Cultural Adaptation (Appendix 14)

Appendix 1

PROMIS® Instrument Maturity Model

Approved: April 11, 2012;
Revised 02/13, 04/13, 05/13

The **Instrument Maturity Model** describes the stages of instrument scientific development from conceptualization through evidence of psychometric properties in multiple diverse populations. The model is used in conjunction with the standards and guidance documents (http://www.nihpromis.org/science/publications?AspxAutoDetectCookieSupport=1) to assist developers in meeting the progressive scientific standard criteria from item pool or scale development to fully validated instruments ready for use in clinical research and practice.

Brief descriptions of each stage follows:

Stage 1: Developmental – Conceptualization & Item Pool Development

The latent trait or domain is conceptualized and defined according to the PROMIS domain framework. Literature reviews and qualitative methods (e.g., individual interviews and/or focus groups) have been used to conceptualize and define the domain. During this phase, attention to literacy, translatability, cultural and lifespan harmonization, and PROMIS guidelines for item construction is required. At the end of this phase, an item pool or scale will have been developed.

Stage 2: Developmental – Calibration Phase

The items have undergone calibration following psychometric analyses using “best practices” factor analysis and item response theory methods or methods appropriate for a different measurement model. In addition, limited information relating the item bank’s measurement properties to existing “legacy” instruments of the domain (concurrent validity) has been assessed. Some modifications to the item pool based on both the qualitative (e.g., cognitive testing or debriefing) and psychometric analyses have been completed. Information has been developed on measurement error across the domain. Instruments such as short forms or CATs have been assessed and defined. Differential item functioning (DIF) is assessed with respect to a minimal set of relevant demographic and language variables (e.g., age, gender, and race/ethnicity), and recommendations made concerning the potential impact of DIF on the use of the item bank and scores. Not all measures will be computer adaptive assessments based on item banks. At times, static forms are desirable or even more appropriate. For example, standardized, static health profile instruments can capture multi-dimensional

health concepts across several item banks. Stage 2 instruments may be appropriate for use as outcome measures in selected research.

Stage 3: Public Release – Calibrated and Preliminary Validation Completed

The measurement properties, validity and reliability of the item bank and related instruments have been more fully assessed and meet the standards for release for public use. A Stage 3 bank meets the same criteria as a Stage 2A or 2B bank for the first eight rows of the Maturity Model. A Stage 3A bank has undergone additional prospective validity and reliability testing than that completed in prior levels. This work may be focused on comparison to an expanded set of legacy measures, which may include a specific clinical population or populations using cross-sectional studies to assess construct validity. The relevance of item content is also further supported in a Stage 3A bank. Stage 3B banks expand the evidence base as relevant to different audiences and applications. For example, they include longitudinal studies to assess responsiveness, mode studies, evaluation of translation into an alternative language and provide some interpretation guidelines in either a general or a clinical population, or both. Targeted data collection facilitates further evaluations for DIF with respect to other covariates beyond those assessed in Stage 2, which now may include education level, socioeconomic status, language translations etc. These item banks and related instruments may be appropriate as clinical research outcomes.

Stage 4: Maturing - Responsiveness and Expansion

These instruments benefit from continued expansion of the development and evaluation begun in Stage 3B. They have undergone continued reliability, validity and responsiveness testing in different clinical populations. DIF analyses have been expanded to include additional relevant known groups which may include socioeconomic status (SES), language translation(s), and literacy levels. These are considered more mature instruments. Based on their measurement characteristics (responsiveness, MIDs, etc.), use within clinical settings (e.g., to measure individual change) may be appropriate. In addition, the underlying item banks may be in the process of being iteratively improved.

Stage 5: Fully Mature User Support

These instruments have undergone very extensive reliability, validity and responsiveness testing across multiple clinical populations. Score interpretations (absolute level or change) have been developed and are used to understand the health of patients and to guide decision making and follow-up actions. These interpretations may emerge as the result of a history of widespread use of the instruments across populations and applications; or, they can be fostered by the developers of the measures who create a user-friendly administration, scoring and interpretation manual or course geared to different audiences for different uses of the measures. The highly mature measure has been widely adopted and used as evidenced by searches in data bases such as PubMed and ClinicalTrials.gov.

These measures have received recognition or endorsement by a formal review process (e.g. COSMIN criteria; Medical Outcomes Trust criteria; FDA qualification, EMA labeling claim review, NQF endorsement, inclusion in DSM, etc.).

	Develop-mental Stage 1A	Develop-mental Stage 1B	Develop-mental Stage 2A	Develop-mental Stage 2B	Public Release in PROMIS/ Assessment Center 3A	Public Release in PROMIS/ Assessment Center 3B	Public Release in PROMIS/ Assessment Center 4	Public Release in PROMIS/ Assessment Center 5
Stage	Item Pool	Prelimi-nary Item Bank	Calibrated Item Bank	Item Bank, Profile or Global Health Measure - Preliminary Reliability/ Validity	Instruments - Validated	Instruments – longitudinal data to for prelim responsiveness – other research to expand use-fullness	Maturing Instruments &/or Item Bank Expansion	Instruments with Fully Mature User Support:
Descriptions	Conceptualized	Ready for Calibration	Dimension ality Assessed & Calibrated	Validity (Construct & Concurrent) – limited	Validity - concurrent & construct validity – cross sectional assessed	Prelim responsiveness	Extensive validity & responsiveness in general and pertinent population samples Item bank modifications - population specific or expansion/ refinement	How scores can be used to understand and respond to health care needs and differences in health is determined & documented
QUALITATIVE: Conceptual documentation and evidence supporting content validity	YES	YES	YES	YES	YES	YES	YES	YES
Dimensionality Specified	NO	YES	YES	YES	YES	YES	YES	YES
Domain Placement Specified (approved)	NO	YES	YES	YES	YES	YES	YES	YES
Item response theory (IRT): Item calibration; information and DIF analyses	NO	NO	YES	YES	YES	YES	YES	YES
Classical test theory (CTT): Evidence supporting dimensionality, reliability and validity (e.g. concurrent validity with legacy)	NO	NO	YES	YES	YES	YES	YES	YES
DIF Preliminary Assessed in Known Groups (e.g. age, race/ethnicity, and gender)	NO	NO	YES	YES	YES	YES	YES	YES

POPULATION: Sample variability reflects variability in construct	NO	NO	YES	YES	YES	YES	YES	YES
FORMAT: CAT and short form measures; Computer, paper forms	NO	NO	YES	YES	YES	YES	YES	YES
Scoring Algorithm Specified	NO	NO	NO	YES	YES	YES	YES	YES
Continued Documentation of Relevance of Item Content and Generalizability as needed	NO	NO	NO	NO	YES	YES	YES	YES
Validity: Concurrent and construct assessed with legacy measures	NO	NO	NO	NO	YES	YES	YES	YES
POPULATION: Expanded DIF analyses relevant population characteristics (e.g. educational status, socioeconomic status etc.)	NO	NO	NO	NO	YES	YES	YES	YES
CTT: Evidence supporting responsiveness and interpretation guidelines (MID, responder criteria)	NO	NO	NO	NO	NO	YES	YES	YES
POPULATION: Translation into one language that is spoken by large percentage of population (e.g. in US, Spanish languages.)	NO	NO	NO	NO	NO	YES	YES	YES
POPULATION: Evaluation in general population and multiple disease conditions including DIF analyses by health condition and language translations.	NO	NO	NO	NO	NO	YES	YES	YES
MODE: Evidence supporting multiple modes of administration (CAT, paper, IVRS, computer)	NO	NO	NO	NO	NO	NO	YES	YES
Continued expansion of DIF analyses across subpopulations as well as continued qualitative work on content validity and to generate items at the tails of the distribution.	NO	NO	NO	NO	NO	NO	YES	YES
POPULATION: Translation and psychometric evaluation into languages other than English	NO	NO	NO	NO	NO	NO	NO	YES
Measure is recognized/certified/endorsed /qualified by a recognized consensus review process conducted by NQF or FDA, for example.	NO	NO	NO	NO	NO	NO	NO	YES

Appendix 2. PROMIS GUIDELINE DOCUMENT	
TOPIC: Domain Framework and Definition	
Authored By: William Riley	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

Scope: This guidance pertains to the processes involved in domain conceptualization, definitions, domain structure, as well as consideration for existing domain structure. Item bank merging and reconciliation are also addressed in this document.

Suggested Developmental Processes:

1. Initial Working Definitions and Domain Framework Location:
   - Devised based on existing literature review, both theoretical and empirical
   - Augmented by analyses of existing data sets when available (archival analyes)
   - Developed consistent with proposed or probable use of the bank/scale

2. Revision of Initial Working Definition based on Expert Review
   - Obtain feedback on working definition from content experts
     - Consider a range of experts (e.g. scale development, outcomes researchers)
     - Independent of research team
     - Sufficient N to achieve saturation (typically 5 – 10)
     - Modified Delphi procedure recommended but other procedures, such as semi-structured interviews, can be used
   - Revise Definition and Framework location based on expert feedback in conjunction with the Domain Framework committee
   - Insure that definition sufficiently bounds or limits the concept and in plain language (no jargon or obscure scientific terminology) to guide patient feedback on item content

3. Revision based on Patient/Respondent Feedback
   - Patients/respondents not expected to provide feedback on the domain definition or framework, but it is possible during focus group procedures for item generation (as described by item bank development committee) that patient feedback may expand or contract, or otherwise shape the domain definition or its position in the framework
   - Document any revisions to the domain definition or framework location based on feedback from patients
   - If substantial revisions are required, repeat step 2 and 3.

4. Revision based on Psychometric Testing
   - Utilize analysis plan to test hypothesized factor structures, subdomains, and item fit within these domains and subdomains
   - Test fit of items with separate but related domains to insure best fit with assigned domain(s)
   - Evaluate relationship of developed domain with existing domains in framework as possible to influence decisions about framework location
   - Items retained for bank should be the result of discussion and compromise between analysts and content experts to select best fit items that also sufficiently address all hypothesized facets of the domain definition – decisions about inclusion and exclusion should be documented.

- If all items representing a facet have a poor fit in the bank and will not be included in the calibrated bank, the domain team should revise and narrow the domain definition to reflect the absence of this facet.
- If items fit poorly with hypothesized domains and subdomains, fit poorly with each other, and/or multiple facets of the hypothesized domain are no longer represented, additional item development and psychometric testing should be considered (i.e. repeat steps 1-4)

5. Revision based on Subsequent Expert Feedback
   - Preferably in all cases, but particularly when most or all items representing a facet of a domain have been pruned, expert feedback on the revised definition should be obtained
   - Expert selection similar to step 2 above (range, independent, sufficient N for saturation)
   - Experts review content of calibrated item bank and provide feedback on how the definition could be revised to better reflect the content of the retained items in the bank
   - Revised domain definition based on expert feedback
   - Obtain Steering Committee review and approval before making definition available to end users.

Item bank merging and reconciliation

Scope:

Item banks and instruments that exist, either internal or external to PROMIS, that cover conceptually similar areas will be reviewed by the domain framework sub-committee, as well as each instrument's developers. Examples of banks that could be considered for reconciliation are: pediatric and adult physical function item banks, or a social role participation bank focused on work roles that complements some of the existing social role banks.

Processes:

The groups for each instrument along with the domain framework sub-committee will discuss, reconcile definitions and domain framework as needed and provide recommendation on to the Steering Committee for vote.

Qualitative and quantitative evidence that support the expansion of existing item banks to include new items should include review of existing domain definitions and reconciliation with development of a new domain definition if the conceptual basis is modified.

Appendix 3. PROMIS GUIDANCE DOCUMENT	
TOPIC: Qualitative Methods	
Authored By: Susan Magasi, PhD, Jennifer Huang, PhD, Liz Jansky, PhD, Margaret Vernon, PhD	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: recommended/common practice	

Study design, including methods of data collection and sampling strategies depend on the purpose of the data, the clinical and demographic characteristics of the population, and the sensitivity of the study topic. Careful planning of the qualitative methods is critical to the quality and validity of the qualitative data.

Design – The research design must be appropriate for the study purpose and population. Qualitative methods are frequently used to gather clinical and content expert input, patient input, and to cognitively evaluate items for comprehension and relevance. Both focus groups and individual interviews can yield valuable data to inform instrument development and refinement.

a. The decision between focus groups and/or individual interviews depends in part on logistical considerations, including: prevalence of the condition; severity of condition; sensitivity of the topic; developmental issues (e.g. susceptibility to peer pressure, group think; and other logistical consideration).

b. Focus group – consensus building, identification of common factors, ideally suited to situations where participants may need to "bounce ideas off each other".

c. Individual Interviews – understand experience in depth; provide a rigorous yet viable alternative when logistical considerations make focus groups impractical or inappropriate.

d. Cognitive Interviews – evaluate if items are easily understood by the target population. Specifically probe comprehension, recall, and response options.

Sampling – Inclusion of a well targeted sample of respondents is critical to the quality and validity of the qualitative data. It is essential that the sample include people with different manifestations of the concept in order to be representative of the experience.

a. Theoretically/conceptually driven to include adequate stratification of the condition/concept across the population.
b. Inclusion and exclusion criteria are clearly documented
c. Sample size not determined a priori but based on data saturation

Data Collection and Interviewing – Data collection is critical to the rigor and validity of the qualitative data. Unlike quantitative methods, that require standardized administration of study materials, qualitative methods also require skilled facilitation and elicitation of data. Facilitator/interviewer training and a well-crafted question route are critical.

a. Documentation of facilitator/interviewer training in qualitative methods. We recommend at a minimum 2 co-facilitators for all focus groups, e.g. 1 lead facilitator and 1 note taker. For individual interviews, we recommend a single interviewer (serving as both facilitator and note taker) with audiotape back-up. Additional note takers may be included but researchers should weigh the value of adding additional research staff with facilitating rapport and participant comfort.
b. Data collection methods are appropriate for the sample – e.g. individual interviews versus focus groups; in-person versus telephone interviews. Need to consider demographic and clinical characteristics of the group.
c. Focus group composition - if using focus groups. - composition must be appropriate (e.g. at least 3 groups, groups sufficiently "focused", group size within accepted guidelines of 6-12 participants [adjusted based on demographic and clinical needs of the participants]).

d. Questioning route/interview guide development – semi-structured, open-ended interview guide that allow for spontaneous responses to emerge. Facilitators should probe participants to gain in-depth information on emergent themes.
e. Data recording and documentation – We recommend audiorecording of all interviews and focus groups, with the option of verbatim transcription (with identifiers removed for analysis) supplemented with detailed structured field notes by facilitators/interviewers.
f. Documentation of compliance with all confidentiality standards as indicated by individual institutional reviews, including de-identification of data, data storage, destruction of recordings, etc.

Analysis – Qualitative data analysis differs from traditional, positivistic research in the integration of data collection and analysis activities, data in the form of text rather than numbers, and the central role that the research team has in the analytic process. Implementation of a systematic approach to qualitative analysis with can help ensure that trustworthiness of the qualitative findings.

a. Documented training of analysis team.
b. All sessions coded by at least 2 coders using a common data dictionary with regular harmonization of newly emergent codes. We recommend double coding of a minimum of 10% of data with regular meetings to confirm reliability.
c. Use of constant comparative methods to identify intra and inter-group differences
d. Analytic strategy that proceeds from descriptive coding (labeling individual comments) focused coding (grouping individual codes into conceptual categories)
e. Analysis and data collection are iterative processes with each process informing the other (e.g. use of emergent themes to flesh out emerging concepts)
f. Documentation of data saturation, e.g. development of a saturation grid.
g. Documentation of **reliability** between coders and within individual coders
h. Triangulation of data from multiple sources, esp. literature review, expert interviews, patient/person interviews
i. Documentation of analytic decisions and how they informed the development of conceptual framework and item content, e.g. the development of an audit trail.

SOFTWARE

Atlas-ti

NVivo

KEY REFERENCES & RESOURCES

Bazeley P. (2007) Qualitative Data Analysis with NVivo. Thousand Oaks, CA: Sage.
Brod M, Tesler LE, Christensen TL. (2009). Qualitative research and content validity: developing best practices based on science and experience. ***Qual Life Res*** 18:1263-1278.

Charmaz K. (2006). Constructing Grounded Theory: A Practical Guide Though Qualitative Analysis. Washington, DC: Sage.
Creswell JW. (2007). Qualitative Inquiry & Research Design: Choosing among Five Approaches. (2nd ed.). Thousand Oaks, CA: Sage.
DeWalt DA, Rothrock, Yount, Stone AA. (2007) PROMIS qualitative item review. ***Medical Care*** 45(5) (Suppl. 1): S12-S21.
Lasch KE, Marquis P, Vigneux M, Abetz L, Arnould B, Bayliss M, Crawford, Rosa K. (2010). PRO development: rigorous qualitative research as the crucial foundation. ***Qual Life Res*** 19:1087-1096.
Leidy NK, Vernon M. (2008) Perspective on patient-reported outcomes: Content validity and qualitative research in a changing clinical trial environment. ***Pharmacoeconomics*** 26(5):363-370.
Merriam SB. (2009) Qualitative Research: A Guide to Design and Implementation. San Francisco, CA: Jossey-Bass.
Miles MB, Huberman AM. (1994) Qualitative Data Analysis: An Expanded Sourcebook. (2nd ed). Thousand Oaks, CA: Sage.
Strauss A, Corbin J. (1997) Grounded Theory in Practice. Thousand Oaks, CA: Sage.

Appendix 4. PROMIS GUIDELINE DOCUMENT	
TOPIC: Structure, Composition and Item ID Names of Individual Items	
Authored By: Susan Magasi, Nan Rothrock	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

SCOPE/ SYNOPSIS

The focus of this document is on the composition of individual items – context, stem, responses. Recommended practices are provided based on PROMIS 1 and 2 experiences. Guidelines for naming the items in a manner consistent with PROMIS items are provided to facilitate the transition to publically available PROMIS item banks (in Assessment Center). Items that vary from these guidance points may be acceptable, and should provide a rationale for the variance.

KEY REFERENCES & RESOURCES

DeWalt DA, Rothrock N, Yount S, Stone AA; PROMIS Cooperative Group. Evaluation of item candidates: the PROMIS qualitative item review. Med Care. 2007 May;45(5 Suppl 1):S12-21.

PROCESSES & RECOMMENDATIONS

a. A comprehensive review of the biomedical and social sciences peer-reviewed literature will be completed to identify reported previously developed instruments that are operational measures of the target concept.
b. Items should measure the target concept, be written at an appropriate reading level for the intended respondents, be clear, concise, sensitive to choice of tense and person, and be grammatically correct.
c. Each multiple choice item includes a context (e.g. setting or timeframe), stem and response. Responses should be grammatically consistent with the stem and be of parallel structure and of approximately the same length. Existing PROMIS response options should be used, whenever possible, or justification for other response options should be provided.
d. Items will be worded to support translation to other languages and for use within multi-cultural contexts.

1. Key considerations for assessing item surface characteristics
   - Item content is consistent with the domain definition
   - Clarity
     - Avoid vague, confusing, long, and complex questions
     - Avoid use of slang
     - [Item uses words and language commonly used by children 8 and older]
   - Precision
     - Avoid “double-barreled” item or multiple examples in item
   - General applicability
     - Avoid item content too narrow to have universal applicability, including stem of the item being disease specific
   - Acceptability to patients
   - Adaptation to computerized format (stand alone on pc screen)
   - Images – recommend use only in combination with other formats that are accessible to low vision or easily amenable to screen readers and other assistive technologies

   Other major reasons for elimination of item

- Item is semantically redundant with a previous item
- Concerns about translatability

2. Response options
   - The PROMIS consensus process acknowledged the need for some uniformity in response options. Given the lack of empirical evidence that one set is clearly better than others, they recommended that one of the preferred response options be used when possible. Most of the PROMIS response option categories include two preferred sets. The majority of PROMIS items used these options with the flexibility to use a different set if an item could not be satisfactorily reworded to fit one of the preferred sets. (For example, pain intensity items are traditionally scored on a 0 to 10 point scale.)
   - The optimal number of response levels may vary for individual items, latent constructs, and context of item administration.
   - Use “not applicable” response options with care and only if deemed necessary.

Category	**Preferred Option Response Set**	**Preferred Option Response Set**
Frequency	*Never*	*Never*
	Rarely	*Once a week or less*
	Sometimes	*Once every few days*
	Often	*Once a day*
	Always	*Every few hours*
Duration	*A few minutes*	*None*
	Several minutes to an hour	*1 day*
	Several hours	*2–3 days*
	1–2 days	*4–5 days*
	>2 days	*6–7 days*
Intensity	*None*	*Not at all*
	Mild	*A little bit*
	Moderate	*Somewhat*
	Severe	*Quite a bit*
	Very severe	*Very much*
Capability	*Without any difficulty*	
	With a little difficulty	
	With some difficulty	
	With much difficulty	
	Unable to do	

3. Recall
   - PROMIS investigators were concerned about selecting a recall period that would reduce the potential biases and yet be sufficient to capture a period of experience that was considered clinically relevant for outcome research. Relatively little research is available to inform this question, but their guiding principle was that relatively shorter reporting periods were to be preferred over longer ones to generate the most accurate data. A 7-day reporting period was adopted as a general convention for most PROMIS items.

- One PROMIS domain, physical function, chose to not specify a time period, but to ask the question in the present tense (e.g. “Currently, do you…”)
- In PROMIS I, Stone et al. conducted some work that aimed to test the accuracy of different recall periods. We are following up as to whether there is a summary of these findings.

4. Literacy level analysis
    - While literacy level requirements were not implemented in PROMIS I, investigators made a substantial effort to create and use items that were accessible in terms of literacy level and that had little ambiguity or cognitive difficulty. All writers targeted the sixth-grade reading level or less, although this proved to be more difficult with some constructs (e.g. social constructs requiring phrases indicating a situation or specific activity and then an assessment of satisfaction about participation versus declarative statements about mood). Writers also attempted to choose words used commonly in English, and avoided idiomatic examples or slang.
    - For items with specific quantities (e.g. 5 pounds, 1 mile) – include both the English and metric values.
    - All items selected for cognitive testing were subjected to testing with the Lexile Analyzer to assess readability. The Lexile Analyzer gives an approximate reading level for the item based on the commonness of words in the item and the complexity of the syntax. Some are not in favor of Lexile analyses as they are intended for passages of text, not single sentences.

5. Cognitive interviewing
    - Minimum of 5 participants reviewing each item
    - If, after 5 interviews the item underwent major revisions, the item was subjected to 3 to 5 additional interviews after the revisions
    - At least 1 nonwhite interviewee and at least 1 white interviewee
    - At least 2 interviewees with one or more of the following criteria:
        - less than 12 years of education;
        - a measured reading level less than the ninth grade using the Wide Range Achievement Test-3 Reading subtest; or
        - a diagnosis associated with cognitive impairment (e.g. traumatic brain injury or stroke).
    - **Lessons Learned from Cognitive Interviewing** Source: *The Life Story of a PROMIS Depression Item: 18 Steps to Validity, Paul A. Pilkonis, PhD, Department of Psychiatry University of Pittsburgh Medical Center*
        - Differing interpretations of time frame
            - Disregarding time frame
            - Focus on sentinel events
            - Past week not equal to past 7 days
        - Many of the problem items had double negatives E.g. “I could not control my temper” and response option of “never”
        - Specifiers can limit generalizability of item – Remove “normal” or “daily” from “household chores”
        - Unintended interpretations E.g. “I felt tingling” was interpreted as a sexual phenomenon by 3/5 respondents and item was eliminated from anxiety pool

6. Miscellaneous

- In PROMIS1, there are banks that utilize “I” and some that use “You”. In either case, uniformity within a given bank or a set of related banks is recommended. In addition, the first-person subject is generally preferred.

Response Options for PROMIS

The following response options were selected by the PROMIS network for use in the development of the initial item pools. These options were finalized 1/30/06. The response options used by the final version 1.0 item banks are listed.

Response Options	Used in Version 1.0 Adult Bank
Frequency #1 Never Rarely Sometimes Often Always	Anger (all except 1 item) Anxiety (entire bank) Depression (entire bank) Fatigue (part of bank) Pain Impact (part of bank) Sleep Disturbance (part of bank) Wake Disturbance (part of bank) Used in modified format by Pain Behavior (entire bank)
Frequency #2 Never Once a week or less Once every few days Once a day Every few hours	Pain Impact (1 item only)
Duration #1 A few minutes Several minutes to an hour Several hours A day or two More than 2 days	Not used by any bank
Duration #2 None 1 day 2-3 days 4-5 days 6-7 days	Fatigue (1 item only)
Intensity #1 (severity) None Mild Moderate Severe Very Severe	Fatigue (1 item only)

Response Options	Used in Version 1.0 Adult Bank
Intensity #2 (or interference) Not at all A little bit Somewhat Quite a bit Very much	Anger (1 item only) Fatigue (part of bank) Pain Impact (part of bank) Sat. with Discretionary Social Activities (entire bank) Sat. with Social Roles (entire bank) Sleep Disturbance (part of bank) Wake Disturbance (part of bank)
Difficulty Without difficulty With some difficulty With much difficulty Unable to do	Used in modified format by Physical Function (part of bank)

MODIFICATIONS BY DOMAIN GROUP

Some domain groups made revisions to the existing response options.

Physical Functioning

- "Difficulty" rating modified as:
  - Without any difficulty
  - With a little difficulty
  - With some difficulty
  - With much difficulty
  - Unable to do
- "Intensity 2" modified as:
  - Not at all
  - Very little
  - Somewhat
  - Quite a lot
  - Cannot do
- For one item, created an additional Difficulty scale:
  - No difficulty at all
  - A little bit of difficulty
  - Some difficulty
  - A lot of difficulty
  - Can't do because of health

Pain Behavior

- Frequency #1 modified as:
  - Had no pain
  - Never
  - Rarely
  - Sometimes
  - Often
  - Always

Sleep Disturbance

- For one item, created an additional Intensity scale:
  - Very poor

- Poor
- Fair
- Good
- Very good

PROMIS Item Naming Conventions

PROMIS has been developing and adapting item naming conventions to aid in communication about individual items. This section describes the current conventions to be used for naming PROMIS items that is needed prior to calibration testing or loading into Assessment Center.

The consistent naming of items serves three purposes: 1) an item's domain can be quickly identified by knowing its item ID 2) the writing of scoring or other analytic scripts will be facilitated as the IDs are more meaningful and 3) the IDs do not imply any unintended intellectual property status associated with a legacy instrument.

The following guidelines should be adhered to as able when naming PROMIS items:

- Eight (8) character limit if possible,
- Alphanumeric characters only
- Do not mix upper and lower case letters in a variable name.
- First 3 – 5 characters should be derived from the domain name (e.g. PAININ, EDANX, GLOBAL)
- Last 2 – 3 characters is a number that is typically based on sequence in calibration testing
- Leave room for growth in the numbers (e.g. use "001" rather than "1")
- Due to variable naming restrictions in SAS and some of the other tools used for data collection by panel companies, we suggest not beginning an item ID with a number and avoiding all special characters (including underscores)

Please note that an item ID only represents one combination of context, stem and response options. An existing PROMIS stem ID **cannot** be utilized for another unique item.

- *

Appendix 5. PROMIS GUIDELINE DOCUMENT	
TOPIC: Translatability & Cultural Harmonization Review	
Written By: Helena Correia	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

SCOPE/ SYNOPSIS

PROMIS items are intended to be appropriate for culturally diverse populations and for multilingual translation. Conducting a translatability review during the item development phase is a standard procedure for PROMIS instruments. This assessment may result in the identification of potential conceptual or linguistic difficulties in specific wording and lead to item revisions. Reviewers may offer alternative wording solutions more suitable for a culturally diverse population, for translation, and for the survey's mode of administration.

This document describes the standard method and criteria for assessing translatability of each individual item and response set. The criteria outlined below reflect the most common issues found during review of PRO instruments in general and during the PROMIS v1 review process in particular. However, they are not static or limiting criteria. Depending on the nature of the subject or domain, the target population, or the type of survey administration, other issues might be noted.

PROCESSES

Overview

The classification outlined below was used in PROMIS 1 and recently revised to include additional categories as well as explanations and examples for each category. The number next to each category is simply an ID or code for that category. Those numbers do not represent a rating of importance or incidence of the issue.

Most of the issues identified through these categories are pertinent in the context of translation into other languages. In addition, the resolution of some of these issues is

also relevant for improving the English version. Ultimately, the translatability review helps to clarify the intended meaning of each item.

An item can have more than one type of issue. The reviewer should list and comment on all aspects that s/he finds problematic. Each reviewer relies on personal experience with translation and knowledge of a particular language besides English, to inform the review comments, with the understanding that no translatability review can cover all possible translation difficulties for all languages.

Categories for classification of issues:

1 = No apparent translatability issues – the reviewer does not foresee a problem conveying the meaning of the item in other languages, and cannot think of any reason why the item should be revised or avoided.

2 = Double negative - negative wording in the item may create a double negative with the negative end of the rating scale for that item ("never" OR "not at all"), making it difficult to select an answer. The negative wording can be explicit (e.g. "I do not have energy") or implicit (e.g. "I lack energy"). There may not be an equivalent implicit negative in other languages.

3 = Idiomatic, colloquial, or jargon – the item contains metaphorical expressions or uses words/phrases in a way that is peculiar or characteristic of a particular language (idiomatic); is too familiar or appropriate only for informal conversation (colloquial); or uses specialized language concerned with a particular subject, culture, or profession, or street-talk (jargon). Other languages may not have equivalent expressions (e.g. "feeling blue", "flying off the handle").

4 = Cultural relevance - the item is not relevant outside of the US or is relevant only to subcultures within the US. For example, the word "block" may not have the same meaning in other countries and therefore not be relevant to measure distance. Also, items containing examples of physical or social activities that are region-specific (e.g. "skiing") or common among specific socio-economic strata (e.g. "playing polo", or playing golf").

5 = Confusing syntax - if the reviewer has to read the item more than once to understand what it says, it's probably not phrased in a simple and clear way.

6 = Split context from item stem - the item is an incomplete sentence or question following an introductory statement or context that begins at the top of page. Not all languages will be able to part the sentence in the same place or at all.

7 = Grammatical structure - sentence structure, use of gerunds or tenses that do not translate well because they do not exist in other languages (e.g. present perfect does not exist in all languages).

8 = Units of measure are country-specific – the item contains references to units of measure (e.g. miles, pounds, yards, etc.) not used outside of the US, or not well understood by people living in the US but who grew up in other countries. *[Reviewer can propose a universal version if possible.]*

9 = Vague, ambiguous or unclear – the item or words in the item could mean different things to different people. For example the word “cool” can be understood as “good” or “interesting”, instead of related to cold temperature, especially by children.

10 = Word that does not translate well – this category covers difficulties in translation that should be noted for the purpose of agreeing on acceptable alternative wording solutions for the translations (how far can the translations go and still harmonize with the source language?). For example, "My child looked upset" may have to be translated as "My child appeared to be upset" or "My child’s face looked upset" or even "My child’s face/my child’s expression gave [others] the impression that he/she was upset".

11 = Elevated or high register language – the item contains words that are too difficult for the average person to understand or are not appropriate for the specific target group (e.g. “thrashing” or “searing” used in a pediatric instrument), or are too technical (e.g. “orthodontic”).

12 = Double barreled item - two or more different terms or concepts are used in the same item, making it difficult to answer the item if only one of the concepts applies to the respondent (e.g. “I lost or gained weight”).

13 = Possible redundancy between items – the item is too similar to another item in the same bank and translation might have to be the same for both (e.g. "I stayed away from other people" and “I kept to myself”; “I felt exhausted” and “I felt wiped out”).

14 = Redundant terms used in the same item – other languages may not have the same variety of terms reflecting different nuances of the same concept as in English (e.g. “aches and pains,” “hurting pain”).

0 = Legacy item – the item is from a validated instrument, and it is being used exactly as it was developed originally (same context, same item stem, same answer

categories), and therefore cannot be revised. *[An independent reviewer may or may not be aware of legacy items.]*

Documenting the translatability review

All reviewers' comments and suggestions are presented or compiled in the same spreadsheet containing the items. If possible, reviewers should start each comment with a key word/expression (e.g. "idiomatic", "redundant", "double-barreled") to facilitate post-review decisions. They must explain what the issue is and if possible offer a suggestion for revision where applicable. Comments should be consistent for the same issues.

When to do the translatability review

The translatability review should be performed once the items are considered almost final, but while changes can still be implemented.

Ideally, and timeline permitting, the translatability review would be done before the cognitive interviews, so that feedback from the reviewer(s) can be incorporated into the interview script. For example if an item is considered ambiguous for translation purposes, it would be beneficial to probe English speakers to find out how they understand that item. As another example, if two items would overlap in meaning after translation, it might be preferable to re-write one of them and debrief the revised version, or debrief both items in the cognitive interviews to assess which one to keep. The assessment of translatability does not focus only on whether an item can be translated into other languages as is. The review can also provide useful input for refining the English version as well.

Implementing the recommendations made by the reviewer(s)

There is no hard rule here. The developers take all the qualitative information into consideration to make a decision on the final wording of each item.

KEY REFERENCES & RESOURCES

Correia, H (2010). PROMIS Statistical Center. Northwestern University. PROMIS Translation Methodology- The Minimum Standard. PowerPoint Presentation.

Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof 2005;28(2):212-32.

Kim J, Keininger DL, Becker S, Crawley JA. Simultaneous development of the Pediatric GERD Caregiver Impact Questionnaire (PGCIQ) in American English and American Spanish. Health and Quality of Life Outcomes 2005; 3:5

Appendix 6. PROMIS GUIDELINE DOCUMENT	
TOPIC: Literacy Review	
Written By: N. Rothrock	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

SCOPE/ SYNOPSIS

PROMIS items are intended to be appropriate for use across a broad range of individuals. As such, items should have the lowest demand on reading skills as possible while retaining the item's meaning. Therefore, during item development items should be reviewed to assess their literacy level. If possible, items for adults should be at a 6th grade reading level or lower.

PROCESSES

Overview

All items should be reviewed to assess their reading level. Items are assessed individually, not as a full item bank/short form.

Specifics *(a table/checklist format suggested)*

Different assessment tools can be used to assess reading level of an item (e.g., Lexile). If an item intended for adults is above a 6th grade reading level, the study team should evaluate how it can be simplified and retain its meaning. Generally readability is improved with shorter sentences and use of frequently used words. For example, "depression" has a lower reading level than "dysthymia".

OVERVIEW of PROMIS –CURRENT PRACTICES

All item stems are reviewed by the Lexile Analyzer (http://lexile.com/analyzer/). Study teams improve readability through revision if possible.

EMERGING ISSUES and FUTURE DIRECTIONS

Most literacy evaluation tools are maximized for passages of text, not single statements.

Appendix 7. PROMIS GUIDELINE DOCUMENT	
TOPIC: Intellectual Property	
Written By: N. Rothrock & A. Stone	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

SCOPE:

This standard describes the process to clarify intellectual property rights of PROMIS measures.

SYNOPSIS:

PROMIS instruments were developed with the intent of making them freely available to clinical researchers. Items from existing instruments required permission from the instrument author for inclusion in PROMIS with the understanding that 1) PROMIS would label all measures as © PROMIS Health Organization and PROMIS Cooperative Group; 2) PROMIS would not collect royalties on behalf of itself or any other investigator; 3) all publications and presentations of results from studies using these instruments should include a statement that PROMIS version x instruments were used; and 4) permission to use PROMIS instruments does not include permission to modify wording or layout of items, distribute to others for a fee, or translate items into another language.

KEY CONCEPTS & DEFINITIONS:

Intellectual Property: distinct creations of an individual(s) for which a set of exclusive rights are granted to the owner.

PROCESSES

Overview

All PROMIS items are owned and controlled by the PROMIS Health Organization and PROMIS Cooperative Group. Items adopted from other instruments into

PROMIS should have explicit permission for use and agreement to the PROMIS Terms and Conditions of Use prior to inclusion.

Specifics *(a table/checklist format suggested)*

1) Items created de novo by PROMIS investigators are owned by the PROMIS Health Organization and PROMIS Cooperative Group. An individual PROMIS investigator cannot claim exclusive intellectual property rights to any item or instrument.
2) Items that are modifications from existing measures have been reviewed to determine if another author has a reasonable claim of intellectual property. If so, that author is requested to provide written permission for use of a modification of his/her item(s) in PROMIS adhering to the PROMIS Terms and Conditions of Use.
3) Items that are adopted as-is by PROMIS require written permission for inclusion in PROMIS and agreement to the PROMIS Terms and Conditions.
4) Authors that provide written permission for use of their items in PROMIS will be included in the PROMIS Terms and Conditions for Use table of contributors.

OVERVIEW of PROMIS –CURRENT PRACTICES *(table format suggested)*

Items are reviewed to identify if they are a clear derivative of a single item that is owned by another investigator. This review includes assessment of the level of similarity between the PROMIS and non-PROMIS item and the number of similar items that exist in other measures. The review includes the item context, stem, and responses.

COMPARISON OF METHODS AND RECOMMENDATIONS *(given the minimum criteria, offer recommendations or guidelines to achieve the next level)*

All PROMIS items require intellectual property rights be assigned to the PROMIS Health Organization and PROMIS Cooperative Group.

EMERGING ISSUES and FUTURE DIRECTIONS

As the exact model for the public/private partnership develops, the role of the PROMIS Health Organization and PROMIS Cooperative Group as intellectual property owners may be modified.

Appendix 8. PROMIS GUIDELINE DOCUMENT	
TOPIC: **Measurement Model**	
Written By: Dennis Revicki & Carole Tucker	
Approved By: SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

SCOPE

Describes the steps and processes involved in calibrating an item bank.

DEFINITIONS & KEY CONCEPTS

Unidimensionality: One critical assumption of IRT models relates to the unidimensionality of the set of items, that is, the items represent a single underlying construct. No item set will ever perfectly meet strictly defined unidimensionality assumptions.[1] The objective is to assess whether scales are "essentially" or "sufficiently" unidimensional[2] to allow unbiased scaling of individuals on a common latent trait. One important criterion is the robustness of item parameter estimates, which can be examined by removing items that may represent a significant dimension. If the item parameters (in particular the item discrimination parameters or factor loadings) significantly change, then this may indicate insufficient unidimensionality.[3,4] A number of researchers have recommended methods and considerations for evaluating essential unidimensionality.[1,2,5-7]

Local Independence: Local independence assumes that once the dominant factor influencing a person's response to an item is controlled, there should be no significant association among item responses.[21-23] The existence of local dependencies that influence IRT parameter estimates represent a potential problem for scale construction or CAT implementation and require additional handling during instrument specification. Scoring respondents based on miss-specified models will result in inaccurate estimates of their level on the underlying trait. Uncontrolled local dependence (LD) among items in a CAT assessment could result in a score less or un- related to the PRO construct being measured.

PROCESSES

Traditional Descriptive Statistics
• Item Analysis:
➢ Response frequency, mean, standard deviation, range, skewness and kurtosis
➢ Inter-item correlation matrix, item-scale correlations, drop in coefficient alpha
• Scale Analysis:
➢ Mean, standard deviation, range, skewness, kurtosis, internal consistency reliability (coefficient alpha)
Evaluate Assumptions of the Item Response Theory Model
• Unidimensionality
➢ Confirmatory Factor Analysis (CFA) using polychoric correlations (one-factor and bi-factor models)
➢ Exploratory Factor Analysis will be performed if CFA shows poor fit.
• Local Independence
➢ Examine residual correlation matrix after first factor removed in factor analysis.
➢ IRT based tests of local dependence
• Monotonicity
➢ Graph item mean scores conditional on total score minus item score.
➢ Examine initial probability functions from non-parametric IRT models
Fit Item Response Theory (IRT) Model to Data
• Estimate IRT model parameters
➢ Samejima's Graded Response Model for unidimensional polytomous response data
• Examine model fit
➢ Compare observed and expected response frequencies
➢ Examine fit indices: S-X^2, Bock's χ^2, and Q_1 statistics
• Evaluate item properties
➢ IRT category response curves
➢ IRT item information curves
• Evaluate scale properties
➢ IRT scale information function
Evaluate Differential Item Functioning (DIF) among key demographic and clinical groups
• Qualitative analyses and generation of DIF hypotheses
• Evaluation of presence and impact of DIF using IRT-based methods:
➢ General IRT-based likelihood-ratio test
➢ Raju's signed and unsigned area tests and DFIT methods
• Evaluation of presence and impact of DIF using Non IRT-based methods
➢ Ordinal logistic regression
➢ Multi-group multiple-indicator, multiple cause (MIMC) model using structural equation modeling
Item Calibration for Item Banking
• Design for administration of PROMIS items for calibration phase
➢ Full bank and incomplete block designs for administration of items to respondents for each item pool. *See Sampling standards document for recommendations*

• Standardize theta metric
➢ Standardizing metric so that general US population has a mean of zero and standard deviation of one. All disease/disorder groups will have a population mean and standard deviation ratio relative to this reference group.
• Assign item properties for each item in the bank.
➢ Calibrate each item with a discrimination parameter and threshold parameters using Samejima's Graded Response Model.
➢ Design or specify parameters for CAT algorithms.

SPECIFICS

Classical Test Theory Methods to Assess Unidimensionality: Prior to assessing dimensionality, several basic classical test theory statistics will be estimated in order to provide descriptive information about the performance of the item set. These include inter-item correlations, item-scale correlations, and internal consistency reliability. Cronbach's coefficient alpha[8] will be used to examine internal consistency with 0.70 to 0.80 as an accepted minimum for group level measurement and 0.90 to 0.95 as an accepted minimum for individual level measurement.

Factor Analysis Methods to Assess Unidimensionality

Confirmatory factor analysis (CFA) should be performed to evaluate the extent that the item pool measures a dominant trait that is consistent with the content experts' definition of the domain. CFA was selected as the first step because each potential pool of items were carefully developed to represent a dominant construct based on an exhaustive literature review and qualitative research.[9] Because of the ordinal nature of the patient-reported outcome (PRO) data, appropriate software (e.g., MPLUS[10] or LISREL[11]) should be used to evaluate polychoric correlations using an appropriate estimator (e.g., the weighted least squares with adjustments for the mean and variance (WLSMV[12] in MPLUS[10]) estimator or diagonally weighted least squares (DWLS in LISREL[11]) estimator for factor analysis.

CFA model fit should be assessed by examining multiple indices. Noting that statistical criteria like the chi-square statistic are sensitive to sample size, a range of practical fit indices should be examined such as the comparative fit index (CFI > 0.95 for excellent fit), root mean square error of approximation (RMSEA < 0.06 for good fit), Tucker-Lewis Index (TLI > 0.95 for excellent fit), standardized root mean residuals (SRMR < 0.08 for good fit), and average absolute residual correlations (< 0.10 for good fit).[2,13-17]

If the CFA shows poor fit, exploratory factor analysis should be conducted and the magnitude of eigenvalues for the larger factors examined (at least 20% of the variability on the first factor is especially desirable), as well as differences in the magnitude of eigenvalues between factors (a ratio in excess of four is supportive of the unidimensionality assumption), scree test, parallel analysis, correlations among factors, and factor loadings to determine the underlying structural patterns.

An alternate method to determine whether the items are "sufficiently" unidimensional is McDonald's bi-factor model[2] (see also Gibbons[18,19]). McDonald's approach to assessing

unidimensionality is to assign each item to a specific sub-domain based on theoretical considerations. A model is then fit with each item loading on a common factor and on a specific sub-domain (group factor). The common factor is defined by all the items, while each sub-domain is defined by several items in the pool. The factors are constrained to be mutually uncorrelated so that all covariance is partitioned either into loadings on the common factor or onto the sub-domain factors. If the standardized loadings on the common factor are all salient (defined as >0.30) and substantially larger than loadings on the group factors, the item pool is thought to be "sufficiently homogeneous".[2] Further, one can compare individual scores under a bi-factor and unidimensional model. If scores are highly correlated (e.g., $r > 0.90$), this is further evidence that the effects of multidimensionality is ignorable.[20]

Local Independence

Local independence assumes that once the dominant factor influencing a person's response to an item is controlled, there should be no significant association among item responses.[21-23] The existence of local dependencies that influence IRT parameter estimates represent a problem for scale construction or single-factor model CAT implementation. Scoring respondents based on miss-specified models will result in inaccurate estimates of their level on the underlying trait. Uncontrolled local dependence (LD) among items in a CAT assessment could result in a score unrelated to the PRO construct being measured.

Identification of LD among polytomous response items includes examining the residual correlation matrix produced by the single factor CFA. High residual correlations (greater than 0.2) should be flagged and considered as possible LD. In addition, IRT-based tests of LD should be utilized including Yen's Q3 statistic[24] and Chen and Thissen's LD indices.[25] These statistics are based on a process that involves fitting a unidimensional IRT model to the data, and then examining the residual covariation between pairs of items, which should be zero if the unidimensional model fits.

The modification indices (MIs) of structural equation modeling (SEM) software may also serve as statistics to detect LD. When inter-item polychoric correlations are fitted with a one-factor model, the result is a limited information parameter estimation scheme for the graded normal ogive model. The MIs for such a model are one degree of freedom chi-square scaled statistics that suggest un-modeled excess covariation between items, which in the context of item factor analysis, is indicative of LD.

Items that are flagged as LD should be examined to evaluate their effect on IRT parameter estimates. One test is to remove one of the items with LD, and to examine changes in IRT model parameter estimates and in factor loadings for all other items.

One solution to control the influence of LD on item and person parameter estimates is omitting one of the items with LD. If this is not feasible because both items provide a substantial amount of information, then LD items can be marked as "enemies," preventing them from both being administered in a single assessment to any individual. The LD need to be controlled in the calibration step to remove the influence of the highly correlated items. In all cases, the LD items should be evaluated to understand the source of the dependency. Another possible option for a pair of items with LD called item A and item B would be to calibrate the scale without item A to

obtain item parameters for item B, and then calibrate the scale again without item B to obtain item parameters for item A. In this way, the influence of LD on the rest of the scale is omitted, but both items A and B are included in the item bank. This permits the inclusion of all of the items without distorting any particular item's information content.

Monotonicity

The assumption of monotonicity means that the probability of endorsing or selecting an item response indicative of better health status should increase as the underlying level of health increases. This is a basic requirement for IRT models for items with ordered response categories. Approaches for evaluating monotonicity include examining graphs of item mean scores conditional on "rest-scores" (i.e., total raw scale score minus the item score) using ProGAMMA's MSP software, or fitting a non-parametric IRT model[26] to the data that yields initial IRT probability curve estimates. A non-parametric IRT model fits trace lines for each response to an item without any a priori specification of the order of the responses. The data analyst then examines the fitted trace lines to determine which response alternatives are (empirically) associated with lower levels of the domain and which are associated with higher levels. The shapes of the trace lines may also indicate other departures from monotonicity, such as bimodality. While non-parametric IRT may not be the most (statistically) efficient way to produce the final item analysis and scores for a scale, it can be very informative about the tenability of the assumptions of parametric IRT. Another possible similar but parametric approach would be to use a multinomial rather than ordinal logistic (i.e. graded response) model. The multinomial model does not assume monotonicity, and facilitates direct assessment of the assumption without the inefficiencies of non-parametric models.

Fit Item Response Theory Model

Once the assumptions have been confirmed, IRT models are fit to the data both for item and scale analysis and for item calibration. IRT refers to a family of models that describe, in probabilistic terms, the relationship between a person's response to a survey question and his or her standing (level) on the PRO latent construct (e.g., pain) that the scale measures.[27,28] For every item in a scale, a set of properties (item parameters) are estimated. The item slope or discrimination parameter describes how well the item performs in the scale in terms of the strength of the relationship between the item and the scale. The item difficulty or threshold parameter(s) identifies the location along the construct's latent continuum where the item best discriminates among individuals. Most of the question response formats in PRO assessment are ordered categorical/ordinal/Likert-type formats, so polytomous models such as the graded response model are fit. In this model, there are multiple threshold parameters. Each threshold parameter indicates the location along the latent continuum where an individual is more likely to endorse the higher as opposed to the lower response category (hence "threshold").

After initial analyses of existing data sets, the PROMIS network evaluated both a general IRT model, Samejima's Graded Response Model[29,30](GRM), and two models based on the Rasch model framework, the Partial Credit Model[31] and the Rating Scale Model.[32,33] Based on these analyses, PROMIS network recommended using the GRM in item bank development work.

The GRM is a very flexible model of the parametric, unidimensional, polytomous-response IRT family of models. Because it allows discrimination to vary item-by-item, it typically fits response data better than a one-parameter model.[28,34] Compared to alternative two-parameter models such as the generalized partial credit model, the model is relatively easy to understand and illustrate to "consumers" and retains its functional form when response categories are merged. The GRM offers a flexible framework for modeling the participant responses to examine item and scale properties, to calibrate the items of the item bank, and to score individual response patterns in the PRO assessment. Other IRT models were fit, as needed, for example for the pain behavior item bank.[35] However, the PROMIS network will examine further the fit and added-value of alternate IRT models using PROMIS data.

The unidimensional GRM is a generalization of the IRT two-parameter logistic model for dichotomous response data. The GRM is based on the logistic function that describes, given the level of the trait being measured, the probability that an item response will be observed in *category k or higher*. For ordered responses $X = k$, $k = 1,2,3, \ldots, m_i$, where response m reflects the highest θ value, this probability is defined[29,30,36] as:

$$P(X_i = k|\theta, b_i, a_i) = \frac{1}{1+\exp\left[-a_i(\theta - b_{i,k-1})\right]} - \frac{1}{1+\exp\left[-a_i(\theta - b_{i,k})\right]}$$

This function models the probability of observing each category as a function of the underlying construct. The subscript on m above indicates that the number of response categories does not need to be equal across items. The discrimination (slope) parameter a_i varies by item i in a scale. The threshold parameters b_{ik} varies within an item with the constraint $b_{k-1} < b_k < b_{k+1}$, and represents the point on the θ axis at which the probability passes 50% that the response is in category k or higher. If a model other than the GRM is used, then there should be strong justification provided for that choice?

IRT model fit should be assessed using a number of indices. Residuals between observed and expected response frequencies by item response category should be compared as will fit for different models based on analyses of the size of the differences (residuals). IRTFIT[37] [1] can be used to assess IRT model fit for each item. IRTFIT computes the extension of S-X^2 and S-G^2 for items with more than two responses.[38,39] These statistics estimate the fit of the item responses to the IRT model, that is, whether the responses follow the pattern predicted by the model. Statistically significant differences indicate poor fit. The S-X^2 (a Pearson X^2 statistic) and S-G^2 (a likelihood ratio G^2 statistic) are fit statistics that use the sum score of all items and compare the predicted and observed response frequencies for each level of the scale sum score. The ultimate issue is to what degree misfit affects model performance in terms of the valid scaling of individual differences.[40]

Once analysts are satisfied with the fit of the IRT model to the response data, attention is shifted to analyzing the item and scale properties of the PROMIS domains. The psychometric properties of the items will be examined by review of their item parameter estimates, item response functions or characteristic response curves(CRCs), and item information curves.[41,42] Information curves indicate the range of theta where an item is best at discriminating among

individuals by increasing the precision of person score estimates. Higher information denotes more precision for measuring a person's trait level. The height of the curves (denoting more information) is a function of the discrimination power (*a* parameter) of the item. The location of the information curves is determined by the threshold (*b*) parameter(s) of the item. Information curves indicate which items are most useful for measuring different levels of the measured construct.

Poorly performing items should be reviewed by content experts before the item bank is established. Misfitting items may be retained or revised when they are identified as clinically relevant and no better-fitting alternative is available. Low discriminating items in the tails of the theta distribution (at low or at high levels of the trait being measured) also may be retained or revised to add information for extreme scores where they would not have been retained in better-populated regions of the continuum.

REFERENCES

1. McDonald RP. The dimensionality of test and items. British Journal of Mathematical and Statistical Psychology. 1981;34:100-117.

2. McDonald RP. Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum; 1999.

3. Drasgow F, Parsons CK. Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement. 1983;7:189-199.

4. Harrison DA. Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics. 1986;11:91-115.

5. Roussos L, Stout W. A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement. 1996;20:355-371.

6. Stout W. A nonparametric approach for assessing latent trait unidimensionality. Psychometrika. 1987;52:589-617.

7. Lai J-S, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Quality of Life Research. 2006.

8. Cronbach LJ Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-334.

9. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Medical Care 2007;45(Suppl 1):S12-S-21.

10. Muthén LK, Muthén BO. Mplus User's Guide. Los Angeles, CA: Muthen & Muthen; 1998.

11. Jöreskog KG, Sörbom D, Du Toit S, Du Toit M. LISREL 8: New Statistical Features. Third printing with revisions. Lincolnwood: Scientific Software International. 2003.

12. Muthén B, du Toit SHC, Spisic D. Robust inference using weighted least squared and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Psychometrika 1997;

13.. Kline RB. Principles and practice of structural equation modeling. New York: Guilford Press; 1998.

14. Bentler P. Comparative fit indices in structural models. Psychological Bulletin. 1990;107:238-246.

15. West SG, Finch JF, Curran PJ. SEM with nonnormal variables. In: Hoyle RH, editor. Structural equation modeling: concepts issues and applications. Thousand Oaks, CA: Sage Publications; 1995:56-75.

16. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6:1-55.

17. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. Testing structural equation models. Newbury Park, CA: Sage Publications; 1993.

18. Gibbons RD, Hedeker DR, Bock RD. Full-information item bi-factor analysis. Psychometrika. 1992;57:423-436.

19. Gibbons RD, Bock RD, Hedeker D, et al. Full-information item bi-factor analysis of graded response data. Applied Psychological Measurement. in press.

20. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. Journal of Personality Assessment. 2005;84:228-238.

21. Steinberg L, Thissen D. Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychological Methods. 1996;1:81-97.

22. Wainer H, Thissen D. How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice. 1996;15:22-29.

23. Yen WM. Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement. 1993;30:187-213.

24. Yen WM. Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement. 1984;8:125-145.

25. Chen W-H, Thissen D. Local dependence indexes for item pairs using item response theory. Educational and Behavioral Statistics. 1997;22:265-289.

26. Ramsay JO. A functional approach to modeling test data. In: van der Linden WJ, Hambleton RK, eds. Handbook of Modern Item Response Theory. New York: Springer; 1997:381-394.

27. Hambleton RK, Swaminathan H, Rogers H. Fundamentals of Item Response Theory. Newbury Park, CA: Sage; 1991.

28. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum; 2000.

29. Samejima F. Estimation of latent ability using a response pattern of graded scores, Psychometrika Monograph. 1969; No. 17.

30. Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, eds. Handbook of Modern Item Response Theory. New York: Springer; 1997:85-100.

31. Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982;47:149-174.

32. Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978;43:561-573.

33. Wright BD, Masters GN. Rating Scale Analysis. Chicago: MESA Press; 1982.

34. Thissen D, Orlando M. Item response theory for items scored in two categories. In: Thissen D, Wainer H, eds. Test Scoring. Mahwah, NJ: Lawrence Erlbaum; 2001:73-140.

35. Revicki DA, Chen WH, Harnam N, et al. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain 2009;146(1-2):158-169.

36. Thissen D, Nelson L, Rosa K, and McLeod LD. Item response theory for items scored in more than two categories. In: Thissen D, Wainer H. eds. Test Scoring. Mahwah, NJ: Lawrence Erlbaum; 2001:141-186.

37. Bjorner JB, Smith KJ, Orlando M, Stone C, Thissen D, Sun X. IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models. Lincoln, RI: QualityMetric Incorporated, 2006

38. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Measure 2000;24:50-64.

39. Orlando M, Thissen D. Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. Appl Psychol Measure 2003;27:289-298.

40. Hambleton RK, Han N. Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In: Lenderking WR, Revicki D, eds. Advances in Health Outcomes Research Methods, Measurement, Statistical Analysis, and Clinical Applications. Washington DC: International Society for Quality of Life Research; 2005:57-78.

41. Reeve BB. Item response theory modeling in health outcomes measurement. Expert Review of Pharmacoeconomics and Outcomes Research. 2003:3;131-145.

42. Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, eds. Assessing Quality of Life in Clinical Trials: Methods of Practice. 2nd Edition. Oxford University Press; 2005:55-73.

Appendix 9. PROMIS GUIDELINE DOCUMENT	
TOPIC: Multidimensional item response theory	
Authored By: I-Chan Huang	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: emerging	

Background

The constructs of patient-reported outcomes (PROs) and quality of life (QOL) are usually multidimensional (e.g., physical, psychological and social domains). However, these domains are measured by specific subscales of a more general construct (i.e., the PRO or QOL). In most cases, these domains are moderately or strongly correlated each other. Whether a person can perform great social functioning is conditioned on his/her physical and psychological status. Unfortunately, when we develop and validate PRO instruments, the methods of unidimensional item response theory (IRT) are dominantly used because the parameter estimation procedures for multidimensional IRT (MIRT) were not fully developed or studied. The unidimensional IRT methods are built on the strong assumptions of unidimensionality and local independence (Lord, 1980).

The application of unidimensional IRT models to the data that are not truly unidimensional has significant implications on the estimations of item parameters and underlying latent scores (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983). Theoretically, if a predominant general factor (i.e., PRO or QOL) exists in the data and specific factors (i.e., physical, psychological and social functioning) beyond the general factor are relatively small, the presence of multidimensionality will not affect the estimations of item parameters and the underlying latent scores. If, however, the data are multidimensional with strong specific factors beyond the general factor, the use of unidimensional methods will lead to estimation of item parameters and the underlying latent scores that are drawn toward the strongest factor in the set of item responses.

The aforementioned second scenario can cause serious distortion of the measurement characteristics of the instruments, especially when the factors of the PROs are highly correlated. A study of Folk & Green (1989) examined the effects of using unidimensional methods on two-dimensional data and found that the estimates of underlying scores were distorted to one or the other of the two domains. They also found that the effect was more significant for adaptive tests because the non-dominant factor will not contribute to the scale score estimation (Folk & Green, 1989). This is in part due to the fact that, in the adaptive tests, the item discrimination parameter estimates were used to select items as well as to estimate the underlying scores.

Other advantages of using MIRT to PROs measurement is that the scale reliability and measurement precision can be maximized since the multidimensional models capture as much

information as possible from different domains. Wang and colleges suggested that, when compared to the unidimensional models, the use of multidimensional models can substantially increase the reliability of all domains in the instrument (Wang, et al. 2004). This study also revealed that only 40% of the items were required when using the multidimensional approach to achieve the same level of measurement precision as the unidimensional approach (Wang, et al. 2004). The use of MIRT can significantly increase the efficacy of computerized adaptive tests because the response of each item will contribute to the estimation of underlying scores of PROs on more than one domains at the same time (Segall DO, 1996).

MIRT model

Several multidimensional techniques are available to handle multidimensional PROs data. These models include (1) non-hierarchical multidimensional model, (2) second-order factor model, and (3) bi-factor model. Figure 1 provides the intuitive frameworks with respect to three multidimensional models.

The non-hierarchical model displays the specific domains of PROs on the same level, but specifically accounts for the relationships among the specific domains by modeling their intercorrelations. Two types of non-hierarchical models are designed for studying multidimensionality: between- and within-models. A between-model allows for the dimensions of HRQOL in an instrument to be correlated with each other, and a within-model allows for items to measure more than one dimension simultaneously.

The second-order factor model is comprised of a higher order general factor (i.e., PRO) and several lower order factors (i.e., specific domains). This higher order factor is hypothesized to account for the intercorrelations among lower order factors. The test of second-order model is similar to that of the non-hierarchical model, with the exception that the covariation link between the specific domains will be modeled as a higher order general factor.

The bi-factor model is comprised of a general factor (i.e., PRO) and several group factors (i.e., specific domains). A general factor represents a common trait of PRO which explains intercorrelations among items due to shared item contents. Group factors capture the item covariation that is independent of the covariation due to the general factor. The bi-factor model can help address issue of local dependence violations by modeling correlated items with a specific domain through a general factor.

Model comparison

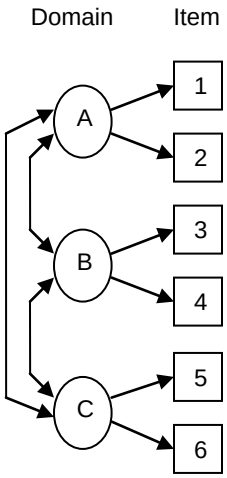
Previous studies have demonstrated the superiority of the bi-factor model to the non-hierarchical multidimensional model and the second-order factor model (Gignac, 2006; Chen, West, & Sousa, 2006; Reise, Morizot, & Hays, 2007). Reise, Morizot, and Hays (2007) reported that the bi-factor model fits the data of quality of provider care better than the non-hierarchical model. Chen, West, and Sousa (2006) reported that the bi-factor model is more appropriate for analyzing the data of mental health than the second-order model. Additionally, the bi-factor method is contains fewer parameters and reduces the model complexity than other competing models (Chen, West & Sousa, 2006). Lai et al., suggest that the bi-factor model can be used to

examine the essential unidimensionality of PROs data (Lai, et al. 2009). Specifically, if the standardized loadings are salient (> 0.3) for all items on the general factor, this suggests that the essential unidimensionality can be held. In contrast, if the loadings of all items on the group factors are salient, this suggests the group factors are well defined and it is more appropriate to report the individual score of the group factors. Reise, Morizot, and Hays argue that when domains are highly correlated to each other (correlation coefficients greater than 0.4), a general factor may exist. In this case, the use of bi-factor model will be an appropriate choice (Reise, Morizot, & Hays, 2007). If, however, the domains are modestly correlated (correlation coefficients between 0.1 and 0.4), the items will tend to have small loadings on the general factor and will have larger loadings on the group factors. In this case, the use of non-hierarchical model will be acceptable (Reise, Morizot & Hays, 2007).

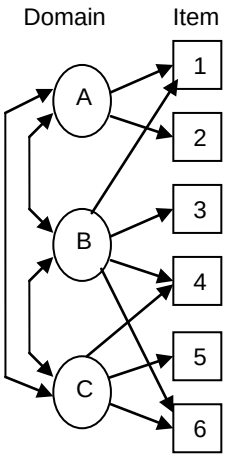
Software

Several analytic models and software can be used to analyze multidimensional data. The measurement model based on a confirmatory factor analysis is a more flexible framework, which allows for conducting the non-hierarchical modeling, second-order factor modeling, and bi-factor modeling. Mplus, for example, is one of the software which can be used to handle multidimensional categorical item response data. Standard fit indexes, such as chi-square index, comparative fit index (CFI), root mean square error of approximation (RMESA), etc. are available to determine the performance of each model. The IRT-based full-information item bi-actor model serves an alternative framework for the bi-factor analysis. This approach is typically based on the marginal maximum likelihood procedure to estimate item parameters. POLYBIF (Gibbons, Bock & Hedeker, 2007) and IRTPRO are among few software for this type of analysis.

Figure 1: Different types of multidimensional modeling for PROs data

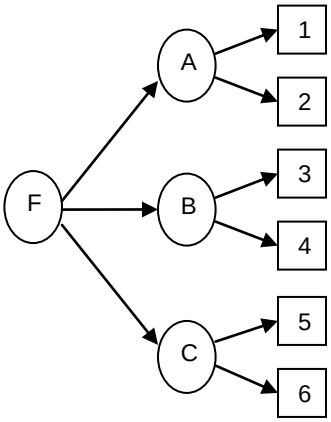


1a. Non-hierarchical model: Between approach

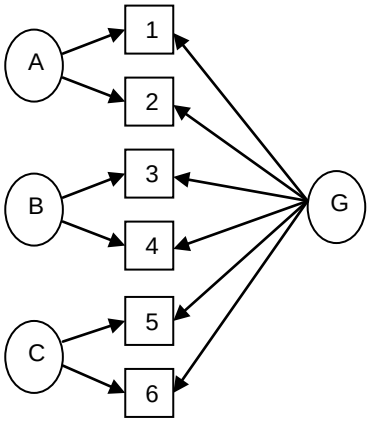


1b. Non-hierarchical model:

Within approach



2. Second-order model



3. Bi-factor model

Note:

Double arrow: domains are correlated with each other; single arrow: domain influences the item response; circle with A, B and C: specific domain; circle with F: a higher-order factor; circle with

References

Ansley TM, Forsyth RA. An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement 1985; 9: 39-48.

Chen FF, West SG, & Sousa KH. A comparison of bifactor and second-order models of quality of life. Multivariate Behavioral Research 2006; 41, 189-225.

Drasgow F, Parsons CK. Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement 1983; 7: 189-199.

Folk VG & Green BF. Adaptive estimation when the unidimensionality assumption of IRT is violated. Applied Psychological Measurement 1989; 13: 373-389.

Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, Kupfer DJ, Frank E, Grochocinski VJ, and Stover A. Full-Information item bifactor analysis of graded response data. Applied Psychological Measurement 2007; 31: 4-19.

Gignac GE. A confirmatory examination of the factor structure of the multidimensional aptitude battery: contrasting oblique, higher order, and nested factor models. Educational and Psychological Measurement 2006; 66: 136-45.

Lai JS, Butt Z, Wagner L, Sweet JJ, Beaumont JL, Vardy J, Jacobsen PB, Shapiro

PJ, Jacobs SR, Cella D. Evaluating the dimensionality of perceived cognitive function. Journal of Pain and Symptom Management 2009; 37: 982-95.

Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum, 1980.

Reise S, Morizot J. and Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research 2007; 16: 19-31.

Segall DO. Multidimensional adaptive testing. Psychometrika 1996; 61: 331-354.

Wang WC, Yao G, Tsai YJ, Wang JD, Hsieh CL. Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. Quality of Life Research 2006; 15:607-20.

Appendix 10. PROMIS GUIDELINE DOCUMENT	
TOPIC: Differential Item Functioning – Identification	
Written By: Jeanne Teresi & Paul Crane, Rich Jones, Jin-shei Lai, Seung Choi, Marjorie Kleinman, Katja Ocepek-Welikson	
Approved By SCC Date: 06/2013	Revision Date: 05/2013
Level: Recommended Processes	

SCOPE:

Description of processes to identify and determine the magnitude and impact of DIF using quantitative methods

SYNOPSIS:

An important goal of the Patient Reported Outcomes Measurement Information System (PROMIS) project is to produce items that can be compared across ethnically diverse groups differing in socio-demographic characteristics. Conceptual and psychometric measurement equivalence of scales are basic requirements for valid cross-cultural and demographic subgroup comparisons. Differential item functioning (DIF) analysis is commonly used to study the performance of items in scales. Different methodologies for detecting DIF in assessment scales have been summarized and compared[i,ii,iii,iv] Item response theory (IRT)[v,vi] and confirmatory factor analysis (CFA)[vii] constitute two general methods of examining item invariance; a discussion of the similarities and differences are summarized in several articles[viii,ix,x,xi,xii;xiii xiv,xv].

KEY CONCEPTS & DEFINITIONS:

DIF: In the context of IRT, DIF is observed when the probability of item response differs across comparison groups such as gender, country or language, after conditioning on (controlling for) level of the state or trait measured, such as depression or physical function.

Uniform DIF: Uniform DIF occurs if the probability of response is consistently higher (or lower) for one of the comparison groups across all levels of the state or trait.

Non-Uniform DIF: Non-uniform DIF is observed when the probability of response is in a different direction for the groups compared at different levels of the state or trait. For example, the response probability might be higher for females than for males at higher levels of the measure of the depression state, and lower for females than for males at lower levels of depression.

Magnitude: The magnitude of DIF relates to the degree of DIF present in an item. In the context of IRT, a measure of magnitude is non-compensatory DIF (NCDIF).[xvi] This index reflects the group difference in expected item scores (EIS). An EIS is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible item categories. Used by Wainer, Sireci and Thissen (1991)[xvii], this effect size measure is frequently used for DIF magnitude assessment. (See also [xviii xix xx xxi xxii xxiii]). Other magnitude measures used in DIF detection include the adjusted odds ratio (logistic regression) or changes in Beta coefficients (hybrid ordinal logistic regression introduced by Crane and colleagues).(See also[xxiv]).

Impact: Expected Scale Score and Differential Test Functioning) Impact refers to the influence of DIF on the scale score. There are various approaches to examining impact, depending on the DIF detection method. In the context of item response theory log likelihood ratio test (IRTLR) results, differences in "test" response functions[xxv] can be constructed by summing the expected item scores to obtain an expected scale score. Plots (for each group) of the expected scale score against the measure of the state or trait (e.g., depression) provides a graphic depiction of the difference in the areas between the curves, and shows the relative impact of DIF. The Differential Test Functioning (DTF) index[xxvi] (Raju and colleagues, 1995) is a summary measure of these differences that incorporate such a weight, and reflects the aggregated net impact. The DTF is the sum of the item-level compensatory DIF indices, and as such reflects the results of DIF cancellation. The latest DFIT software has recently been released[xxvii] In MIMIC and MG-CFA methods, impact can be examined by comparing model-based DIF-adjusted mean scores. Other impact measures are described in several articles[xxviii,xxix].

Anchor Items Anchor items are those items found (through an iterative process or prior analyses) to be free of DIF. These items serve to form a conditioning variable used to link groups in the final DIF analyses.

Purification: Purification is the process of iteratively testing items for DIF so that final estimation of the trait can be made after taking this item-level DIF into account. Purification is described in a separate standard document.

PROCESSES

Overview

1. **Identification of DIF hypothesis**
2. **Study design – sampling plan to provide adequate group sizes for DIF analyses of salient sub-groups.**
3. **DIF analyses**

Specific Approaches

IRT log-likelihood ratio (IRTLR) modeling: The IRTLR likelihood ratio tests[xxx,xxxi,xxxii,xxxiii,xxxiv,xxxv] in IRTLRDIF[xxxvi,xxxvii] and MULTILOG[xxxviii,xxxix], were used for DIF detection in PROMIS 1, accompanied by magnitude measures,[xl] such as the non-compensatory DIF (NCDIF) index [xli,xlii].

Scale level impact was assessed using expected scale scores, expressed as group differences in the total test (scale) response functions, which show the extent to which DIF cancels at the scale level (DIF cancellation).

IRTOLR: The method used as the primary method by most PROMIS 1 investigators was logistic regression and ordinal logistic regression (OLR) using an observed conditioning score. A modification, IRTOLR,[xliii,xliv] was used in some analyses. Estimates from a latent variable IRT model, rather than the traditional observed score are used as the conditioning variable; this method incorporates effect sizes into the uniform DIF detection procedure. DIFwithPAR incorporates trait level estimates to be obtained using the graded response model in PARSCALE.xlv The program allows the user to specify the criteria for DIF, e.g., statistical tests of uniform and non-uniform,[xlvi] an effect size modification based on changes in the pseudo-R^2 in nested models,[xlvii] or a change in coefficient criterion for uniform DIF[xlviii]. Purification is performed in an iterative fashion.

MIMIC: The multiple-indicator, multiple causes (MIMIC) model is a special application of the latent trait model (based on covariance structure analyses using the normal distribution function) that allows DIF to be detected across ethnic/racial groups, after controlling for covariates[xlix]. The model is linked to IRT as originally proposed by Birnbaum[l] because the discrimination parameter can be calculated using the factor loadings (lambdas) (see also[li,lii]).

SIBTEST: Other methods were also used in sensitivity analyses to examine DIF in this item set. These other methods include SIBTEST[liii liv] for binary items and Poly-SIBTEST [lv] for polytomous items. SIBTEST is non-parametric, conditioning on the observed rather than latent variable, and does not detect non-uniform DIF.

EVIDENCE: PROMIS –CURRENT PRACTICES

Principal Investigator(s)	**Subgroups**	**Model**	**Programs**	**Recommendations**
Crane, Heidi Crane, Paul Patrick, Donald **University of Washington**	Spanish vs. English language, gender, HIV transmission risk factor, self-reported race, age, education, illicit drug use, and HIV severity	hybrid IRT/ordinal logistic regression, MIMIC	difwithpar, mplusmimic	Probably will use *lordif*
Forrest, Cristopher **Children's Hospital of Philadelphia**	Gender, race/ethnicity, setting (clinical vs. school), chronic disease group, reading ability, and presumed level of latent trait (involved versus healthy)	Not specified	Not specified	
Fries, James **Stanford**	Age, gender, education level	Ordinal logistic regression	Not specified	Recommend *lordif* if only OLR used;

Principal Investigator(s)	Subgroups	Model	Programs	Recommendations
University				recommend a sensitivity analysis method
Pilkonis, Paul **University of Pittsburgh**	Age, race	IRT likelihood ratio test, ordinal logistic regression	IRTLRDIF	Recommend also examining IRTPRO and perhaps *lordif* for sensitivity analyses for OLR
Potosky, Arnold Moinpour, Carol **Georgetown University; Fred Hutchinson Cancer Research Center**	Race/ethnicity, age	IRTLR, Lord's Wald test (refurbished) MG-CFA, MIMIC, IRTOLR	IRTLRDIF, IRTPRO, DFIT (for magnitude measure-NCDIF) MPlus, *lordif*	
Bode, Rita, Hahn, Elizabeth **PROMIS SC** PROMIS Social Health Supplement	Modes of administration**,** by age (< 65 versus 65+), gender, and education level (≤HS Grad/GED versus ≥Some College).	IRTLR, IRTOLR	IRTLRDIF *lordif*	

COMPARISON OF METHODS AND RECOMMENDATIONS:

Recommendation of a "best" method is difficult because there are so many factors that can impact DIF assessment. Simulations inform about what methods perform better under conditions observed in PROMIS, e.g., skew. Most studies have been conducted with binary items rather than with polytomous items such as those used in PROMIS. Moreover, as new methods are developed, the studies lag behind. In 2006 and 2007 PROMIS investigators reviewed the simulation studies extant (see Teresi 2006). An updated review is being prepared that will include the latest summary of simulation studies, such as those of Carol Woods.[lvi,lvii]

Thus, the PROMIS recommendation is to have a primary method, with another method used in sensitivity analyses. IRT-based methods are recommended. Magnitude of DIF should be assessed, together with both aggregate and individual impact. The table provides some guidelines and recommendations. The approach was to accept as valid the method recommended by the investigator, but in that context suggest software that might be used. Sensitivity analyses were recommended.

REFERENCES

Difwithpar and *lordif* :

Choi SW, Gibbons LE, Crane PK. Lordif : An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulation. Journal of Statistical Software, Under review.

Crane PK, Gibbons LE, Jolley L, Van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. Medical Care. 2006;44:S115-S123.
Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual Life Res. 2007;16(Suppl 1):69-84.
Crane PK, Van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004;23:241-256.
Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement. 1990;27:361-370.
Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html. 1999.

DFIT:

Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999;23:309-32.
Morales LS, Flowers C, Gutiérrez P, Kleinman M, Teresi J A. Item and scale differential functioning of the Mini-Mental Status Exam assessed using the DFIT methodology. Medical Care. 2006;44:S143-151.
Oshima TC, Kushubar S, Scott JC, Raju NS. DFIT8 for Window User's Manual: Differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation.
Raju NS, Van Der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995;19:353-368.

IRTLR:

Orlando-Edelen M, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory ad the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. Medical Care. 2006;44:S134-S142.
Thissen D. IRTLRDIF v2.0b; Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Available on Dave Thissen's web page. 2001.
Thissen D. MULTILOG™ User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software, Inc.; 1991.
Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In PW Holland, H Wainer (Eds). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993:123-135.

IRTPRO:

Cai L, duToit, Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago: Scientific Software, Inc.; 2009.
Langer MM. A re-examination of Lord's Wald test for differential item functioning using item response theory and modern error estimation. Dissertation, University of North Carolina at Chapel Hill, 2008.
Thissen D. IRTPRO: Beta Features and Operation, January 2010.

MIMIC and MG-CFA SEM framework:

Cai L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robins-Monro algorithm. Psychometrika. 2010;75:33-57.

Jöreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Society. 1975;10:631-639.
Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. Medical Care. 2006;44:S124-133.
Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49:115-132.
Muthén LK, Muthén BO. Mplus Users Guide. Version 5 edition. Los Angeles, CA: Muthén & Muthén, 1998-2007.
Muthén B, Asparouhov T. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Los Angeles: University of California and Muthén & Muthén; 2002:16.

GENERAL:

Hambleton RK. Good practices for identifying differential item functioning. Medical Care. 2006;44:SS182-S188.
Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993;17:297-334.
Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Medical Care. 2006;44(Suppl. 11):S152-S170.

---

[i]. Holland PW, Wainer H. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

[ii]. Camilli G, Shepard LA. Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications, 1994;4.

[iii]. van de Vijver F, Leung K. Methods and data analysis for cross-cultural research. Thousand Oaks, CA: Sage Publications, 1997.

[iv]. Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Medical Care. 2006;44:S152-170.

[v]. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum, 1980.

[vi]. Lord FM, Novick MR. Statistical theories of mental test scores (with contributions by A Birnbaum). Reading, MA: Addison-Wesley, 1968.

[vii]. Joreskog K, Sorbom D. LISREL8: Analysis of linear structural relationships: Users Reference Guide. Scientific Software International, Inc., 1996.

[viii]. McDonald RP. A basis for multidimensional item response theory. Applied Psychological Measurement. 2000;24:99-114.

[ix]. Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. Organizational Research Methods. 2004;7:361-381.

[x]. Mellenbergh GJ. Generalized linear item response theory. Psychological Bulletin. 1994;115:302-307.

[xi]. Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993;17:297-334.

[xii]. Raju NS, Laffitte LJ, Byrne BM. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. Journal of Applied Psychology. 2002;87:517-528.

[xiii]. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin.

1993;114:552-566.
[xiv]. Takane Y, De Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987;52:393-408.
[xv]. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Medical Care. 2006;44(Suppl. 11):S152-S170.
[xvi]. Raju NS, Van Der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995;19:353-368.
[xvii]. Wainer H, Sireci SG, Thissen D. Differential testlet functioning: Definitions and detection. Journal of Educational Measurement. 1991;28:197-219.
[xviii]. Chang H, Mazzeo J. The unique correspondence of the item response function and item category response functions in polytomously scored item response models. Psychometrika. 1994;39:391-404.
[xix]. Collins WC, Raju NS, Edwards JE. Assessing differential item functioning in a satisfaction scale. Journal of Applied Psychology. 2000;85:451-461.
[xx]. Morales LS, Flowers C, Gutiérrez P, Kleinman M, Teresi J A. Item and scale differential functioning of the Mini-Mental Status Exam assessed using the DFIT methodology. Medical Care. 2006;44:S143-151.
[xxi]. Orlando-Edelen M, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory ad the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. Medical Care. 2006;44:S134-S142.
[xxii]. Steinberg L, Thissen D. Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods. 2006;11:402-415.
[xxiii]. Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Morales LS, Orlando-Edelen M, Cella D. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measure of physical functioning ability and general distress. Quality Life Research. 2007;16:43-68.
[xxiv]. Monahan PO, McHorney CA, Stump TE, Perkins AJ. Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. Journal of Educational and Behavioral Statistics. 2007;32:92-109.
[xxv]. Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading Massachusetts: Addison-Wesley Publishing Co., 1968.
[xxvi]. Raju NS, Van Der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995;19:353-368.
[xxvii]. Oshima TC, Kushubar S, Scott JC, Raju NS. DFIT8 for Window User's Manual: Differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation.
[xxviii]. Stark S, Chernyshenko OS, Drasgow F. Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? Journal of Applied Psychology. 2004;89:497-508.
[xxix]. Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, 44, 93-116.
[xxx]. Kim SH, Cohen AS. Detection of differential item functioning under the graded response model with the likelihood ratio test. Applied Psychological Measurement 1998;22:345-355.

[xxxi]. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In PW Holland, H Wainer (Eds). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993:123-135.
[xxxii]. Cohen AS, Kim SH, Wollack JA. An investigation of the likelihood ratio test for detection of differential item functioning. Applied Psychological Measurement 1996;20:15-26.
[xxxiii]. Thissen D, Steinberg L, Gerard M. Beyond group-mean differences: The concept of item bias. Psychological Bulletin. 1986;99:118-128.
[xxxiv]. Thissen D, Steinberg L, Gerard M. Beyond group-mean differences: The concept of item bias. Psychological Bulletin. 1986;99:118-128.
[xxxv]. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models, In Holland PW and Wainer H eds. Differential Item Functioning, Lawrence Erlbaum, Inc., Hillsdale NJ, 1993, 123-135.
[xxxvi]. Thissen D. MULTILOG™ User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software, Inc.; 1991.
[xxxvii]. Thissen D. IRTLRDIF v2.0b; Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Available on Dave Thissen's web page. 2001.
[xxxviii]. Thissen D. MULTILOG™ User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software, Inc.; 1991.
[xxxix]. Thissen D. IRTLRDIF v2.0b; Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Available on Dave Thissen's web page. 2001.
[xl]. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. Statistics in Medicine. 2000;19:1651-1683.
[xli]. Raju NS, Van Der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995;19:353-368.
[xlii]. Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999;23:309-32.
[xliii]. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004;23:241-256.
[xliv]. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. Medical Care. 2006;44:S115-S123.
xlv. Muraki E, Bock RD. PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (Version 3). Chicago: Scientific Software, 1996.
[xlvi]. Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement. 1990;27:361-370.
[xlvii]. Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html. 1999.
[xlviii]. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004;23:241-256.
[xlix]. Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49:115-132.
[l]. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Co., 1968.
[li]. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the

parameters of item response models. In PW Holland, H Wainer (Eds). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993:123-135.

[lii]. Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. Medical Care. 2006;44:S124-133.

[liii]. Shealy RT, Stout WF. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika. 1993;58:159-194.

[liv]. Shealy RT, Stout WF. An item response theory model for test bias and differential item functioning. In Holland PW, Wainer H, eds. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993;197-239.

[lv]. Chang H, Mazzeo J, Roussos L. Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. Journal of Educational Measurement. 1996;33:333-353.

[lvi]. Woods CM. Evaluation of MIMIC-model methods for DIF testing with comparison of two group analysis. Multivariate Behavioral Research, 2009, 44, 1-27.

[lvii]. Woods, CM. Likelihood-ratio Dif testing: Effects of nonnormality. Applied Psychological Measurement, 2008, 32, 511-526.

Appendix 11. PROMIS GUIDELINE DOCUMENT	
TOPIC: **Differential Item Functioning – Purification**	
Written By: Jeanne Teresi & Paul Crane, Rich Jones, Jin-shei Lai, Seung Choi, Marjorie Kleinman, Katja Ocepek-Welikson	
Approved By SCC Date: 06/2013	Revision Date: 05/2013
Level: Recommendations	

SCOPE:

This standard describes processes to reconcile the elimination of items that exhibit DIF through purification processes

SYNOPSIS:

An important goal of the Patient Reported Outcomes Measurement Information System (PROMIS) project is to produce items that can be compared across ethnically diverse groups differing in socio-demographic characteristics. Conceptual and psychometric measurement equivalence of scales are basic requirements for valid cross-cultural and demographic subgroup comparisons. Differential item functioning (DIF) analysis is commonly used to study the performance of items in scales. Different methodologies for detecting DIF in assessment scales have been summarized and compared[lviii,lix,lx,lxi] Item response theory (IRT)[lxii,lxiii] and CFA[lxiv] constitute two general methods of examining item invariance; a discussion of the similarities and differences are summarized in several articles[lxv,lxvi,lxvii,lxviii,lxix;lxx lxxi,lxxii].

PROMIS currently recommends that items with DIF are removed from the item bank – hence "absolute purification" is the standard.

KEY CONCEPTS & DEFINITIONS:

DIF: In the context of IRT, DIF is observed when the probability of item response differs across comparison groups such as gender, country or language, after conditioning on (controlling for) level of the state or trait measured, such as depression or physical function.

Magnitude: The magnitude of DIF relates to the degree of DIF present in an item.

Impact: Expected Scale Score and Differential Test Functioning: Impact refers to the influence of DIF on the scale score.

Anchor Items Anchor items are those items found (through an iterative process or prior analyses) to be free of DIF. These items serve to form a conditioning variable used to link groups in the final DIF analyses.

Purification: Item sets that are used to construct preliminary estimates of the attribute assessed, e.g., depression include items with DIF. Thus, estimation of a person's standing on the attribute may be incorrect, using this contaminated estimate. Purification is the process of iteratively testing items for DIF, which may be addressed by the possible removal of these items, so that final estimation of the trait can be made after taking this item-level DIF into account. Simulation studies have shown that many methods of DIF detection are adversely affected by lack of purification. Thus, this process should be considered for incorporation for some methods. Individual impact can be assessed through an examination of changes in depression estimates (thetas) with and without adjustment for DIF. The unadjusted thetas are produced from a model with all item parameters set equal for the two groups. The adjusted thetas are produced from a model with parameters that showed DIF based on the IRTLRDIF results estimated separately (freed) for the groups.

PROCESSES

Overview

1. **Determine the magnitude and impact of DIF (see DEV_DIF1 standard)**
2. **Purification**

This area is a work-in-progress. Currently one can remove an item with DIF from the bank or flag it as an enemy item. There is a multiple calibration feature in the current PROMIS software that was designed to handle an item that is shared across projects. There can be separate calibrations for groups, but they would hold for all items. One item with DIF, such as the crying item could not be calibrated separately. In other words, it is not currently possible to use the PROMIS general population calibrations for all items, and separate group calibrations for specific, e.g., gender groups for a specific item.

Subsequent developmental work by Choi and colleagues would focus on the capability to account for DIF using group specific item parameters. Future research should examine the impact of DIF in computer adaptive testing (CAT). Choi and colleagues are examining the potential for a CAT framework that can account for DIF in real time.

REFERENCES

Difwithpar and *lordif* :

Choi SW, Gibbons LE, Crane PK. Lordif : An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulation. Journal of Statistical Software, Under review.

Crane PK, Gibbons LE, Jolley L, Van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. Medical Care. 2006;44:S115-S123.

Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual Life Res. 2007;16(Suppl 1):69-84.

Crane PK, Van Belle G, Larson EB. Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine. 2004;23:241-256.

Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement. 1990;27:361-370.

Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html. 1999.

DFIT:

Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Applied Psychological Measurement. 1999;23:309-32.

Morales LS, Flowers C, Gutiérrez P, Kleinman M, Teresi J A. Item and scale differential functioning of the Mini-Mental Status Exam assessed using the DFIT methodology. Medical Care. 2006;44:S143-151.

Oshima TC, Kushubar S, Scott JC, Raju NS. DFIT8 for Window User's Manual: Differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation.

Raju NS, Van Der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 1995;19:353-368.

IRTLR:

Orlando-Edelen M, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory ad the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. Medical Care. 2006;44:S134-S142.

Thissen D. IRTLRDIF v2.0b; Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Available on Dave Thissen's web page. 2001.

Thissen D. MULTILOG™ User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software, Inc.; 1991.

Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In PW Holland, H Wainer (Eds). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993:123-135.

IRTPRO:

Cai L, duToit, Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago: Scientific Software, Inc.; 2009.

Langer MM. A re-examination of Lord's Wald test for differential item functioning using item response theory and modern error estimation. Dissertation, University of North Carolina at Chapel Hill, 2008.

Thissen D. IRTPRO: Beta Features and Operation, January 2010.

MIMIC and MG-CFA SEM framework:

Cai L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robins-Monro algorithm. Psychometrika. 2010;75:33-57.

Jöreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Society. 1975;10:631-639.

Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. Medical Care. 2006;44:S124-133.

Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49:115-132.

Muthén LK, Muthén BO. Mplus Users Guide. Version 5 edition. Los Angeles, CA: Muthén & Muthén, 1998-2007.

Muthén B, Asparouhov T. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Los Angeles: University of California and Muthén & Muthén; 2002:16.

GENERAL:

Hambleton RK. Good practices for identifying differential item functioning. Medical Care. 2006;44:SS182-S188.

Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993;17:297-334.

Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Medical Care. 2006;44(Suppl. 11):S152-S170.

[lviii]. Holland PW, Wainer H. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

[lix]. Camilli G, Shepard LA. Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications, 1994;4.

[lx]. van de Vijver F, Leung K. Methods and data analysis for cross-cultural research. Thousand Oaks, CA: Sage Publications, 1997.

[lxi]. Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Medical Care. 2006;44:S152-170.

[lxii]. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum, 1980.

[lxiii]. Lord FM, Novick MR. Statistical theories of mental test scores (with contributions by A Birnbaum). Reading, MA: Addison-Wesley, 1968.

[lxiv]. Joreskog K, Sorbom D. LISREL8: Analysis of linear structural relationships: Users Reference Guide. Scientific Software International, Inc., 1996.

[lxv]. McDonald RP. A basis for multidimensional item response theory. Applied Psychological Measurement. 2000;24:99-114.

[lxvi]. Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. Organizational Research Methods. 2004;7:361-381.

[lxvii]. Mellenbergh GJ. Generalized linear item response theory. Psychological Bulletin. 1994;115:302-307.

[lxviii]. Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993;17:297-334.

[lxix]. Raju NS, Laffitte LJ, Byrne BM. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. Journal of Applied Psychology. 2002;87:517-528.

[lxx]. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin. 1993;114:552-566.

[lxxi]. Takane Y, De Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987;52:393-408.

[lxxii]. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Medical Care. 2006;44(Suppl. 11):S152-S170.

Appendix 12. PROMIS GUIDELINE DOCUMENT	
TOPIC: Validity - Responsiveness	
Written By: Ron Hays, Carole Moinpour et al.	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Recommended	

SCOPE/ SYNOPSIS

This standard describes the evaluation of validity of PROMIS measures.

Evaluation of construct validity will be accomplished through the specification and testing of hypotheses about the relationship between the PROMIS domain scores and clinical status and other relevant variables. A variety of statistical techniques can be employed to evaluate these hypotheses including mixed-effects models, regression analysis, and structural equation models. Interpretation of the PROMIS measures will be facilitated by estimation of minimally important differences (MID) and responsiveness to change. Multiple anchor-based methods are used to provide point estimates and confidence intervals around the estimates. For example, one can evaluate associations between changes in PROMIS domain scores and global ratings of change in clinical status by patients and clinicians and changes in clinical variables. Responsiveness can be represented by effect sizes, standardized response means, and the responsiveness index.

There should be a rationale supporting the particular mix of evidence used to document each type of validity for the measure. Adequate attention to content validity and qualitative methods including review and acceptance by individuals similar to those for whom the measure is designed (Magasi et al: Quality of Life Research, 25 Aug 2011) Construct and concurrent validity (including criterion validity where possible) should be addressed relative to a priori hypothesized relationships with related measures such as other measures of the same domain (criterion or convergent validity) or clinical indicators of severity or existing validated instruments of the target concept (e.g., known groups validity). Rationale and support for the choice of criterion measures should be included in presentations of criterion validity data. The description of the methods and sample used to evaluate each aspect of validity should be provided.

Additional methods for documenting a measure's validity include confirmatory factor analyses to support the instrument's hypothesized factor structure. Differential item functioning (DIF) can identify non-hypothesized differences among respondents reporting a similar level of the trait of interest based on covariates such as age, sex, race, education level. DIF analyses help identify items that should not be included in the

measure or for which appropriate accounting needs to be made when assessing people across different demographic groups. The final instrument should be re-reviewed by experts and end-users/individuals to assess consistency with or identify differences between original definitions and final product.

Identifying minimally important differences (MID) is a part of the process of documenting the cross-sectional or longitudinal construct validity of a measure. MIDs help identify meaningful differences when interpreting the results of known groups comparisons or when determining how sensitive a measure is to change (see below). Both cross-sectional and longitudinal anchor variables can be used to classify patients into distinct groups that have clinical meaning and can therefore help identify MIDs for the new measure; distributional methods have been developed to produce effect sizes with guidelines for inferring importance (Yost & Eton, 2005; Yost et al., 2011). An example of a cross-sectional anchor is using either patient-reported or physician-reported performance status to identify clinically important differences in a measure of physical function. An example of a longitudinal anchor is a patient's report of change in the domain of interest since the last time the patient completed the questionnaire about that domain; example response options for global ratings of change are a lot better, a little better, about the same, a little worse, a lot worse. Change in a clinical anchor such as hemoglobin can also be used to gauge clinically important change. It is also the case that there is not one MID for a PRO measure; the MID can vary based on the sample and the measurement context, requiring the exploration of MIDs in various samples as is done in sensitivity analyses (Revicki et al., 2008).

Responsiveness to change is an important property of an instrument and is used to document the validity of that measure (Hays, 1992). In one sense, responsiveness documents longitudinal validity (Terwee et al., 2007). However, more is required to actually support a measure's responsiveness; there must be evidence that the new measure is detecting known changes in the patient's health (Revicki et al., 2008). Longitudinal data comparing a group that is expected to change with a group that is expected to remain stable is one way to document a measure's sensitivity to change, given that change is expected. As discussed above, change in external anchors (those not reported by a patient such as a physician rating or a lab value) or patient global ratings of change are strategies for supporting a measure's responsiveness. It is important to remember that benchmarks for determining individual change are much larger than those identified for group change due to the additional measurement error present in an individual measure at any point in time (Hays et al., 2005; Donaldson, 2008).

KEY CONCEPTS & DEFINITIONS:

Construct validity. Construct validity refers to the extent to which a measure performs in accordance with a-priori hypotheses and therefore measures what it is intended to measure.

Minimally important differences (MID). A difference in scores that is small but large enough to matter.

Responsiveness to change. Refers to the ability of a measure to reflect underlying change.

SOFTWARE

Standard statistical software such as SAS, Stata and SPSS can be used to evaluate construct validity.

OVERVIEW of PROMIS –CURRENT PRACTICES

The PROMIS Wave II studies are employing a variety of approaches that attempt to use external anchors and hypothesized changes in health-related quality of life to assess the construct validity of the PROMIS domains.

COMPARISON OF METHODS AND RECOMMENDATIONS

Responsiveness to change is estimated using change in the numerator and difference indicators of noise in the denominator: effect size (SD at baseline), standardized response mean (SD of change), and responsiveness statistic (SD of change among those deemed to be "stable"). A variety of estimates should be considered.

EMERGING ISSUES and FUTURE DIRECTIONS

The use of so called "distribution-based" estimates of the MID are increasingly recognized as attempts to compare true estimates ("anchor-based") with prior estimates or beliefs about the magnitude of the MID. Responsiveness to change is related to the estimate of the MID but it expands the evaluation of change beyond the focus on small but important differences to encompass all levels of underlying change.

KEY REFERENCES & RESOURCES

Cronbach L J, Meehl PE. (1955). Construct validity in psychological tests. ***Psychological Bulletin***, 52, 281-302.

Donaldson G. (2008). Patient-reported outcomes and the mandate of measurement. Quality of Life Research, 17, 1303-1313.

Guyatt G, Osoba D, Wu AW, Wyrwich KW, Norman GR. (2002). Methods to explain the clinical significance of health status measures. ***Mayo Clinic Proceedings***, 77, 371-383.

Hays R, Hadorn D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. ***Quality of Life Research***, 1, 73-75.

Hays RD, Brodsky M, Johnston MF, Spritzer KL, Hui K-K. (2005). Evaluating the statistical significance of health-related quality-of-life change in individual patients. Evaluation & the Health Professions, 28, 160-171.

Nunnally JC and Bernstein IH (1994). **Psychometric theory. 3rd ed**,. New York, NY: McGraw-Hill, Inc.

Revicki D, Hays RD, Cella D, Sloan J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology, 61, 102-109.

Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34-42.

Yost KJ, Eton DT. (2005) Combining distribution- and anchor-based approaches to determine minimally important differences. The FACIT experience. Evaluation & the Health Professions,

Yost KJ, Eton DT, Garcia SF, Cella D. (2011) Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. Journal of Clinical Epidemiology, 64, 507-516.

Appendix 13. PROMIS GUIDELINE DOCUMENT	
TOPIC: Reliability	
Written By: Ron Hays, Carol Moinpour et al	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Recommended	

SCOPE/ SYNOPSIS

This standard describes the evaluation of reliability of PROMIS measures.

Reliability will be assessed using internal consistency reliability (coefficient alpha), test-retest reliability, and scale information functions. Internal consistency reliability is estimated using a two-fixed effects model partitioning variance between respondents from the interaction between items and respondents. Test-retest reliability can be estimated by the intraclass correlation estimated from a two-way random effects model that partitions between respondent variance from variance due to time of assessment. Information (precision is estimated at different points along the construct continuum. Reliability in its basic formulation is equal to 1- SE^2, where the SE (standard error) = 1/ $(information)^{1/2}$.

KEY CONCEPTS & DEFINITIONS:

Internal consistency reliability: Estimate of the reliability of a multi-item scale that is based on correlations among the items and the number of items in a scale.

Test-retest reliability: Estimate of reliability that evaluates the association between the scale score at two or more time points.

Information: Larger information is associated with higher reliability and lower standard error of measurement. Information is conditional on where one is along the underlying continuum.

Standard error: An estimate of the uncertainty around a point estimate of a score that can be used to set confidence intervals.

PROCESSES

Overview: Reliability can range between 0-1 with higher being better. A reliability of 0.70 is recommended for group comparisons and 0.90 or higher for individual assessment.

Specifics: PROMIS domain scores have been shown routinely to have adequate reliability for group comparisons. For individual-level administration of PROMIS item banks, the conventional default stopping rule is a SE of 0.30 or less (reliability of 0.91).

SOFTWARE

Standard statistical software such as SAS and SPSS can be used to estimate reliability of measures.

OVERVIEW of PROMIS –CURRENT PRACTICES

PROMIS relies upon estimates of information and associated standard errors throughout the theta continuum for domain scores. Internal consistency reliability is not as useful for long item banks and for computer-adaptive testing where different people are administered different items.

KEY REFERENCES & RESOURCES

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, California: Sage.

Nunnally J. (1978). Psychometric theory, 2nd edition. New York: McGraw-Hill.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Appendix 14. PROMIS GUIDELINE DOCUMENT	
TOPIC: Translation and Cultural Adaptation	
Written By: Helena Correia	
Approved By SCC Date: 06/2013	**Revision Date:** 05/2013
Level: Standard	

SCOPE/ SYNOPSIS

PROMIS projects do not have to include translation. However, translation is a likely choice given the recognized need to make the PROMIS item banks accessible to Spanish-speakers (and ideally Chinese-speakers as well) residing in the United States, and researchers' interest in developing culturally appropriate approaches to health status assessment (e.g. comparing across ethnically diverse groups with different socio-demographic characteristics.) There is also growing interest in using PROMIS measures outside of the United States, and this will necessarily involve translation.

This standard describes the methodology for translating PROMIS instruments. Translations result from an iterative process of forward and back-translation, bilingual expert review, and pre-testing with cognitive debriefing (linguistic validation). Harmonization across all languages and a universal approach to translation guide the process.

The translation methodology described below is considered a minimum standard for ensuring accurate and culturally appropriate translations, as well as equivalence between the languages. There can be enhancements to this methodology if resources are available.

KEY CONCEPTS & DEFINITIONS:

Harmonization across languages - Language and culture prevent us from converting an English item into an exact replica in the target language. Not all languages will use the same exact words but they all must convey the same meaning. Some variation in the choice of words across languages is unavoidable. Harmonization is the process of determining the range of acceptable variation.

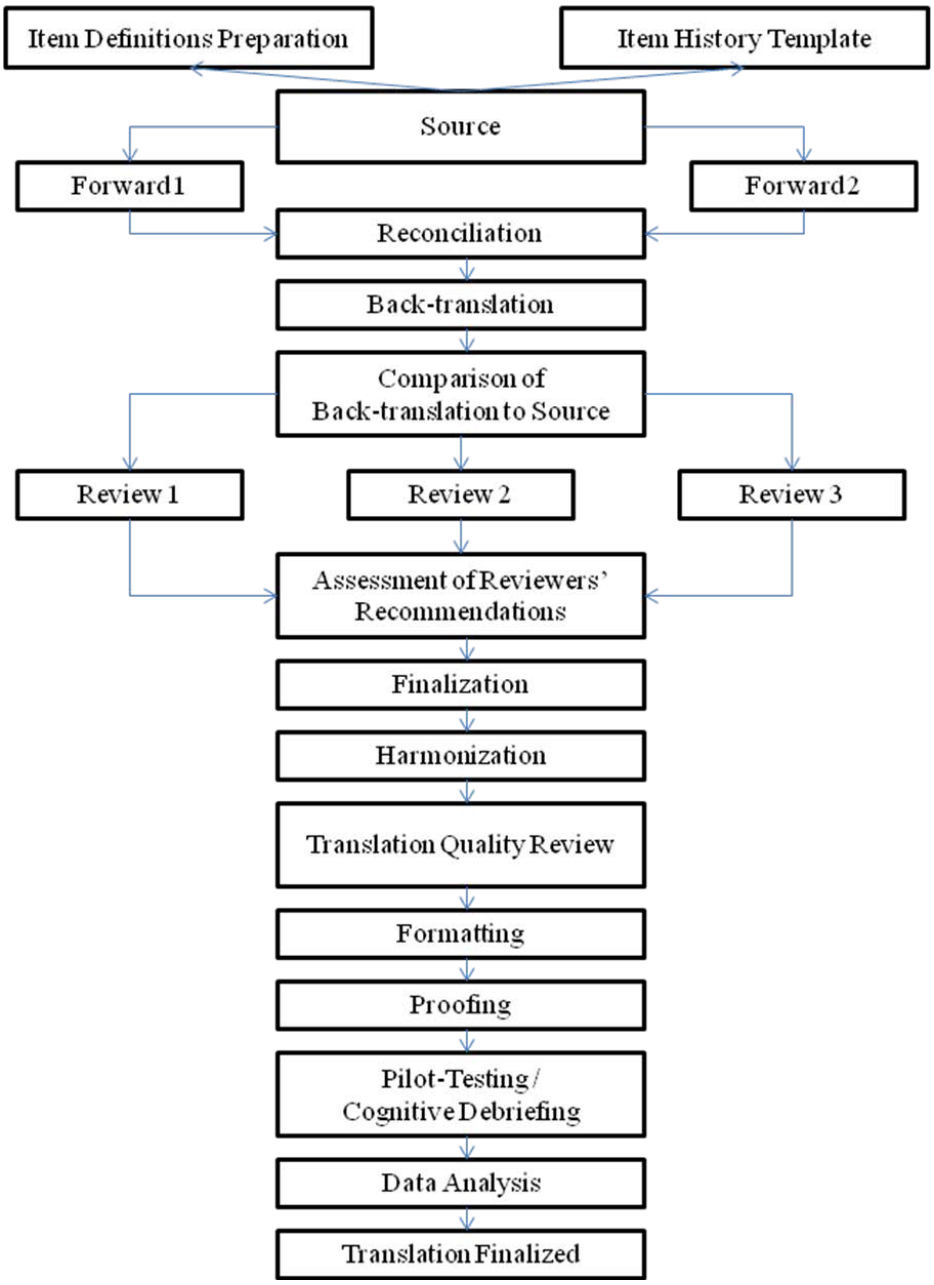
Universal approach to translation – The goal is to create one language version for multiple countries instead of country-specific versions of the same language. Several strategies are employed to reach a universal version: 1) translators from various countries or dialects contribute to the translation process; 2) avoiding colloquial and idiomatic expressions; 3) pretesting and debriefing items with samples from relevant countries.

PROCESSES

Overview

All items, item context(s), and answer options are translated using the FACIT translation methodology (Bonomi, Cella, Hahn, et al., 1996; Eremenco, Cella, Arnold, 2005.) This methodology was employed in the translation of all PROMIS v1 adult and pediatric items and is consistent with the guidelines recommended by the International Society for Pharmacoeconomic and Outcomes Research (ISPOR) for translation of PRO instruments (Wild, Grove, Martin, et al., 2005).

FACIT Translation Methodology - Chart



Specifics *(a table/checklist format suggested)*

The steps of the FACIT translation methodology are described in more detail below:

1) **Two simultaneous forward translations (2 Fwd)**: Source items in English are translated into target language by two independent professional translators who are native speakers of the target language.
2) **Reconciled single target language translation (1 Rec)**: A third independent translator, also a native speaker of the target language, reconciles the two forward translations by selecting one of the forward translations, creating a hybrid version, or providing a new version. Translator must also note the reasons why the reconciled version is the best way to convey the meaning of the source.

3) **Back-translation (1 BT)**: This reconciled version is then back-translated by a native English-speaking translator who is fluent in the target language. The translator does not see the original English source items or item definitions. The back-translation into English must reflect what the target language translation says, without embellishing it.

4) Back-translation review: The Translation Project Manager (TPM) compares source and back-translated English versions to identify discrepancies in the back-translations and to provide clarification to the reviewers on the intent behind the items. This step also results in a preliminary assessment of harmonization between the languages.

5) **Expert reviews (3 Revs)**: Three experts who are native speakers of the target language, independently examine all of the preceding steps and select the most appropriate translation for each item or provide alternate translations if the previous translations are not acceptable. These reviewers are linguists or healthcare professionals (a mixed group is recommended).

6) Pre-finalization review: The Translation Project Manager evaluates the merit of the reviewer's comments, identify potential problems in their recommended translations, and formulate questions and comments to guide the language coordinator for the target language.

7) **Finalization (1 LC):** The Language Coordinator (LC), native of the target language, who worked on the translation development most likely as a reviewer, determines the final translation by reviewing all the information in the Item History and addressing the TPM's comments. Along with the final translation, the LC also provides the respective literal back-translation and polished back-translation for each item. The LC must explain the choice of final translation and offer justification for the decision if the final translation is different from the reconciled version or from what reviewers recommended individually.

8) **Harmonization and quality assurance**: The Translation Project Manager makes a preliminary assessment of the accuracy and equivalence of the final translation by comparing the final back-translations with the source, and verifying that documentation of the decision making process is complete. A quality review* performed by the PROMIS Statistical Center also addresses consistency with previous translations, with other languages if applicable, as well as between the items. The Language Coordinator may be consulted again for additional input.

9) Formatting, typesetting and proofreading of final questionnaire or item forms by two proofreaders working independently, and reconciliation of the proofreading comments.

10) **Cognitive testing and linguistic validation**: The target language version is pre-tested with participants who are native speakers of the target language. The goal is to have each new item debriefed in the target country by at least 5 participants in a cognitive debriefing interview to verify that the meaning of the item is equivalent to the English source after translation.

11) **Analysis of participants' comments and finalization of translation**: The Translation Project Manager compiles participants' comments (back-translated into English) and summarizes the issues. The Language Coordinator (native of the target language) reviews the issues and proposes translation solutions. The TPM verifies that solutions proposed by the LC harmonize with the source and with other languages.

Documenting the translation process (Item History) - Prior to beginning the translation process, the items are incorporated into a document called an Item History in which each item and its subsequent translations and related comments are listed on a separate page (in the case of a Word document) or a separate column (in the case of an Excel document). This format makes it possible to focus on the translation item by item, and provides a convenient format for the translators and reviewers to visually compare the different translations and back-translation and to provide comments on the translation of each item. The finalized translation of each item is subsequently formatted into the layout appropriate to the project for the pre-testing phase and later the format for final distribution.

Item definitions - Also in preparation for the translation, item definitions are created by the Translation Project Manager and reviewed by the item developers or created by the item developers. The document contains the explanation of the concept measured in each item as well as the technical definition of each of the terms used in the item. The purpose of the item definitions is to clarify the intended meaning of each item, thus

ensuring that the meaning is reflected appropriately in the target language. This document is used as a reference by the Translation Project Manager and all the translators involved in the translation development. The item definitions can be included in the Item History next to each item.

Formatting and proofreading - After all translations are completed in the item histories, they are copied and pasted into the Excel file formats provided by the PROMIS team. In order to store the translations and to facilitate the proofreading step, if possible, both the English items and the translations are uploaded into a translation memory. The translated banks are sent to two proofreaders. Once the proofreading issues are resolved, any changes made to the items at proofreading are documented in the Item History, so that the most up-to-date version of the translated item is always recorded there.

Cognitive debriefing – An interview script template is created by the Translation Project Manager and translated into the target language (one forward translation and one proofreading). The cognitive debriefing script covers all or most items, and the questions can be customized for each language, depending on the type of specific issues that surfaced during the translation process. Each item is debriefed with 5 people, for a total of approximately 35 items per subject. All subjects are recruited from the general population. Each subject is asked to first answer the items independently. Completion of the questionnaire is followed by the cognitive debriefing interview. A target language or bilingual interviewer asks the subject a few general questions to elicit feedback on the difficulty of any items or whether any items are offensive or irrelevant, followed by questions regarding item comprehension (i.e. the meaning of specific words in the items, the overall meaning of the item, or why they chose a specific answer). For some items, the subjects are also asked to consider alternative wording for those items.

All the subjects' comments and suggestions regarding each item are compiled into a Pilot Testing Report (PTR) document, and analyzed by the Translation Project Manager to determine if the items were well understood by the target language population. After reviewing their comments and consulting with the Language Coordinator for the target language, revisions can be made to the translation. Final translations are once again proofread to ensure that post-testing revisions are made consistently within the same banks as well as across banks if the revisions to the target language are applicable to all items using the same wording in English.

*A final quality review by the PROMIS Statistical Center is performed again after the cognitive debriefing step is conducted and the translation is finalized, to verify that standards have been met, documentation is complete and to approve the translation.

Note: In very specific cases the resolution of the translation can also result in revision of the English wording (*decentering*).

Once the translation and linguistic validation process have been completed, the items are available for field testing, calibration and psychometric validation.

KEY REFERENCES & RESOURCES

Bonomi AE, Cella DF, Hahn EA et al. Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. Qual Life Res 1996 June;5(3):309-20.

Cella D, Hernandez L, Bonomi AE et al. Spanish language translation and initial validation of the functional assessment of cancer therapy quality-of-life instrument. Med Care 1998 September;36(9):1407-18.

Correia, H (2010). PROMIS Statistical Center. Northwestern University. PROMIS Translation Methodology- The Minimum Standard. PowerPoint Presentation.

Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Eval Health Prof 2005;28(2):212-32.

Lent L, Hahn E, Eremenco S, Webster K, Cella D. Using cross-cultural input to adapt the Functional Assessment of Chronic Illness Therapy (FACIT) scales. Acta Oncol 1999;38(6):695-702.

Wild D, Grove A, Martin ML, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8:94–104

Wild D, Eremenco S, Mear I, Martin M, Houchin C, Gawlicke M, Hareendran A, et al. Multinational trials-recommendation on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: The ISPOR patient reported outcome translation and linguistic validation good practice task force report. Value in Health 2008;12:430-440.