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. Reference citations are provided at the instrument maturity description of measures derived from the PROM

Instrument development and validat some standards, such as those rela pertain only to those item banks and have achieved levels of validity, relia product is a calibrated item bank, he have limited applicability to these se

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

2.3Tj 0 Tc 0 Tw 0.835 0 Td ( )Tj -0.00
T I

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:

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-

despld 4(t)-3(e-3(at)-3.1((p3(s)-1-16.29)-5(one)]Ton)5(s)-1(houl)1(d b)-1(il)0.9(di)1(ded r)-3.1r)-3()-
Infntie,co9(t)-3(y)-1(of)-3(c)-ntite3(at)-3.1(5(s)-1(houl)1(d be1r)2(ov)-1(i)0.)-1568.-should204w0iET 7

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.

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		

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						

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.

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

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)

o56.

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 p(ns)-1eSmisP1teatethePnss

- 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.
    -

- 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 -0.00(ed by)3.7(anye)-5.6(ban

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

- o Poor
- o Fair
- o Good
- o 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

6 = Split context from item stem - the item is an incomplete sentence or question

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 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

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 couldes of thei T amO it2tocal22 Tm]T

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

• 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- 75 T

unidimensionality is to assign each item to a specific sub-domain based on theoretical considerations. A model is then fit w-(7ISm)3th(gnc8(s)-1.7(h i)-.8(em)3.odel)oa.1(di)-1ng(ed 7( )5a [(c)-1.8(em -c an71(o)5 [(uni)-1.1(di)-1.1(m)3.4(ens)-(7ISm)3o.1(nal)-1.3(m)3.4(odel).7(If 1(ns)-1c8(c)-1.heor)0..1(s)-1.ahe

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

M86g0non-
isespoliemt395(iit. 0.7(eas)apen

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 $X = k$, $k = 1,2,3, ..., m_i$, where response m reflects the highest θ value, this probability is defined[29,30,36] as:

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.

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

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.

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)

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.

Figure 1: Different types of multidimensional modeling for PROs data

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.

Appendix 10. PROMIS

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].473 >>BDC 2 4

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

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*	

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.

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:

Appendix 11. PROMIS GUIDELINE DOCUMENT

TOPIC: **D**

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)

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: S170.

[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.

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 h05 ((w)-4(-3(her(-5.1(l)4ew)-3(D)1e)-1(om)-3(pal)0.w -35.85(he)5(n

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

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

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

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

;aM-o5(SM-1(aM-17w

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 Rev)**: These experts are native speakers of the target language, independently examine all of the preceding steps and select the most

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

ci

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