PROMIS®

Minimum requirements for the release of PROMIS instruments after translation and recommendations for further psychometric evaluation

Version: March 3, 2014

SUMMARY

The minimum requirements for the release of short forms and item banks/CATs are described in stage 1. These requirements should be met before PROMIS instruments can be released in a new language. Stage 2 describes further psychometric evaluation of short forms or item banks/CATs that is recommended after the release of short forms and item banks/CATs in a new language. Stage 2 also includes recommendations about how to obtain reference scores in the new language/country.

Stage 1

Instrument	Minimum requirement BEFORE release	Sample size
Short form	Availability of good validation data previous to the translation and a **good quality translation**	Not applicable
Item bank/CAT	Preliminary release: Availability of good validation data previous to the translation and a **good quality translation** Full release: an evaluation of **Differential Item Functioning** (DIF) between language groups and within relevant sub-groups	 200 subjects per group

Stage 2

Instrument	Recommendations for further psychometric evaluation AFTER release	Sample size
Item bank	**Calibration of item banks** in relevant patient groups and the general population;	Minimum 500 Optimal 1000-2000
Short form / item bank	The collection of language- or **country-specific reference scores** in the general population;	At least 300-400
Short form / item bank	An evaluation of the **relevance and comprehensiveness of the items** (content validity) of the item bank and cultural adaptation;	4-6 focus groups or 12 interviews
Short form / CAT	Further psychometric evaluation (**construct validity, internal consistency, test-retest reliability, measurement error, DIF among different patient groups**) of PROMIS instruments in specific patient populations;	200 subjects per group for DIF 50-100 for other measurement properties
Short form / CAT	**Responsiveness and Minimal Important Change** of PROMIS instruments in relevant patient populations.	50-100

Stage 1. Minimum requirements for release

1.1 Short forms

For short forms the minimum requirement for release after translation is the **availability of good validation data previous to the translation** (for example in the US population) and **a good quality translation** (as described in Appendix 14 of the PROMIS standards [1]). Calibration and further psychometric evaluation of the translated short forms (see stage 2) are recommended over time. Evaluation of the short-form calibrations will be strengthened by including questions from the corresponding item bank which were not included in the short form and which are randomly selected from among those not included. For this purpose, 10 or more questions would be desirable.

1.2 Item banks and CATs

For full item banks and CATs the minimum requirement for release after translation is the **availability of good validation data previous to the translation** (for example in the US population) and **a good quality translation** (as described in Appendix 14 of the PROMIS standards [1]). In addition, DIF testing is considered important. A two-stage release procedure was adopted:

1. Preliminary release: After translation, the item bank will be preliminary released under the condition that users share their data with a (local or US) PROMIS group for validation purposes. As soon as data from at least 200 people are available (from one or a combination of multiple studies), **DIF analyses** will be performed and published.
2. Full release: If no major problems are found with DIF testing, the item bank will be fully released. If in stage 1 major DIF problems are found, the translation may be modified (leading to a new version of the item bank), or country-specific item calibrations may be used (leading to a new software version).

Recommendations for the design and analyses of a study on DIF testing are described below.

DIF analyses

Differential Item Functioning (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 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 is observed when the probability of response is in a different direction for the groups compared at different levels of the state or trait.
The aim of this study is to examine whether language-specific item calibrations are needed for some items. The basic assumption of PROMIS is that common item calibrations (at the moment, the US calibration is considered the global calibration database) can be used in all countries, unless it is shown that language-specific ones are needed.

Design: Cross-sectional study

Study population: The study population should be a relevant population for the item bank (e.g. chronic pain patients for the pain item banks) or a sample from the general population of native speakers or individuals whose primary language is the language in which the items are administered.

Sample size: A minimum of 200 subjects per group is recommended for DIF analyses [2].

Data collection: All respondents should **complete all items** of an item bank. If multiple item banks are calibrated, one may need multiple study populations. PROMIS investigators recommend **a maximum of 150 items be administered to a given respondent** (including other items, see below). In addition to the item banks, it is recommended to collect variables to identify appropriate subgroup for testing DIF. This would require, for example, including in the 150 items questions to ascertain the respondents' age, gender, and educational attainment, as well as to collect relevant characteristics of patients, if applicable, for descriptive purposes or known groups analysis.

Additional data requirements: Include variables from a similar study population as that used in the US if it is desired to conduct DIF analyses between language groups (new language compared to US data).

Analyses: At the moment, the US calibration is considered the global calibration database, against which translations should be compared. Different methodologies for detecting DIF in assessment scales have been used. Item Response Theory (IRT) and Confirmatory Factor Analysis (CFA) are two general methods of examining item invariance. In the context of IRT, a measure of magnitude of DIF is non-compensatory DIF. This index reflects the group difference in expected item scores. Also, ordinal regression methods can be used. A change in McFadden's R^2 or adjusted odds ratio may be used to estimate the amount of DIF [3]. Recommendation of a "best" method is difficult because there are many factors that can impact DIF assessment. Thus, the PROMIS recommendation is to have a primary method, with another method used in sensitivity analyses. **IRT-based methods are recommended**. For further information, see the PROMIS standards document (Appendix 10) [1]. It is encouraged to test not only for DIF between language groups but also for DIF among sub-groups (age, sex, education) within the new language.

Criteria for DIF: A change of 2% in McFadden's R^2 has been suggested as a criterion for DIF [3]. It is also important to assess the impact of DIF on the scale score. There are various approaches to examining impact, depending on the DIF detection method. For example, plots of theta against the total item score for all items and for the items with DIF, show the relative impact of DIF. If DIF among language groups is found, it is recommended to simulate the impact of DIF on CAT-based theta estimates (see e.g. [4]).

If DIF among language groups is found, country-specific item calibrations may need to be used in CATs.

Minimum documentation required: the following information should be documented:

- statistical method used (e.g. ordinal logistic regression), criterion used for DIF (e.g. McFadden's R^2 change of 2%), names of the items that showed non-uniform and uniform DIF (e.g. PFB5). Plots, showing the relationship of Theta and the total item score for all items and for the items with DIF are recommended.

Stage 2. Recommendations for further psychometric evaluation and references scores

The PROMIS investigators recommend further psychometric evaluation after the release of PROMIS instruments in a new language or country. These analyses need not require primary data collection and can be conducted, for example, using PROMIS data collected in prior studies where the measures were fielded. There recommendations include:

1. **Calibration of item banks** in relevant patient groups and the general population;
2. The collection of language- or **country-specific reference scores** in the general population;
3. An evaluation of the **relevance and comprehensiveness of the items** (content validity) of the item bank and cultural adaptation;
4. Further psychometric evaluation (**construct validity, internal consistency, test-retest reliability, measurement error, DIF among different patient groups**) of PROMIS instruments in specific patient populations;
5. **Responsiveness and Minimal Important Change** of PROMIS instruments in relevant patient populations.

2.1 Calibration of item banks

PROMIS recommends to perform full calibration analyses on a data set that includes the relevant patient populations as well as in the general population of the country in which the PROMIS instruments have been released. PROMIS recommends to use common item calibrations, except for items with DIF, where language-specific calibrations can be applied when those items are used. This enables comparisons across languages, and data pooling for clinical trials.

Design: Cross-sectional study

Study population: The study population should be a relevant population for the item bank (e.g. chronic pain patients for the pain item banks) or a sample from the general population. Most important in the selection of the study population is **that there should be enough variation in the construct** being measured among the subjects to enable calibration of all items. One should aim to get a heterogeneous sample with regard to what is being measured, including people at the extremes. **A pilot study is recommended** wherein data are collected by at least 30 but ideally 100 subjects to get a sense of the endorsement of the response levels.

Sample size: Different recommendations have been found in the literature for estimating item parameters using a 2-parameter model, ranging from 250 to 2000 patients [5, 6].
Reise and Yu concluded that at least 500 subjects are needed to achieve an adequate calibration under the graded response model. However, for good estimations of the easiest and most difficult items, they recommend 2000 subjects [6]. **PROMIS recommends a minimum of 500 subjects per item** (i.e. each item should have been completed by at least 500 subjects). It should be noted that this sample size may be adequate for estimating item parameters, but may be too small for other analyses, such as computing item and test information functions. Also inflated discrimination parameters can be a problem. Therefore, a more optimal sample size would be 1000 to 2000 subjects per item. When using polytomous models estimating the thresholds at the extremes is challenging when there aren’t enough people responding with the extreme option. It is recommended to have at least 5-10 people with extreme responses on the very easy of difficult items.

It should also be kept in mind that the quality of estimates of parameters is always affected by characteristics of the sample from which they are computed. Characteristics to consider include variation in the construct being measured and respondent motivation to complete a large number of items. The consequences of any mismatch between the study population and the population in which the item bank will be used in the future are unknown.

Data collection: For calibration purposes, all respondents should complete all items. If multiple item banks are calibrated, one may need multiple study populations or use a block design where each subject completes parts of different items banks (see for example [7]). PROMIS investigators recommend **a maximum of 150 items be administered to a given respondent**. This 150 will include non-PROMIS items such as those required to evaluate DIF and, potentially, to support other analyses as explained previously and also below.

In addition to the item banks, it is recommended to **collect relevant variables for testing DIF**, such as age, gender, and race, and to collect relevant characteristics of patients, if applicable, for descriptive purposes. It is also recommended (but not obligatory) to **include a legacy instrument** that measures the same construct as the item bank for initial construct validity testing (e.g. the SF-36 subscale physical functioning or a disease-specific physical functioning subscale for comparison with the PROMIS item bank physical functioning).

Additional data requirements: Including subject characteristic variables that were used in similar study populations as those used in the US is recommended for DIF analyses between language groups (new language compared to US data).

Analyses: The analyses should be performed **as described in the PROMIS standards document** [1]. Such standards include traditional descriptive statistics of items and scale, assumptions of the IRT model (unidimensionality, local independence, monotonicity), IRT model fit, and item calibrations.

It is encouraged (but not required) to test for DIF among demographic groups (age, sex, race), and for DIF among language groups, if possible. At the moment, the US calibration is considered the global calibration database, against which translations should be compared. If DIF among language groups is found, it is recommended to simulate the impact of DIF on CAT-based theta estimates (see e.g. [4]). If possible, a correlation between the theta score on the PROMIS item bank with the legacy instrument should be calculated.

Criteria for good calibration: In the PROMIS standards document (Appendix 8) criteria are provided for what constitutes good calibration results [1]. These criteria should be used as guidelines, not strict cut-off values. If the translated items bank does not meet these criteria, the local and US PROMIS groups should discuss whether release of the items should be postponed of whether the item bank may need to be modified.

Minimum documentation required: To evaluate the minimum standards the following information should be documented:

- **Dimensionality**: results from Exploratory Factor Analyses (percentage of variance explained by the first factor) or Confirmatory Factor Analyses (fit statistics, such as Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR));
- **Local independence**: residual correlations among items, or results from IRT-based tests;
- **IRT calibration and fit**: IRT model used (e.g. Samejima's Graded Response Model for unidimensional polytomous response data), software package used (e.g. Multilog), method of estimation (e.g. maximum likelihood or Bayesian estimation method), model fit (e.g. s-X2,

s-G2, Bock's χ2, Q1 statistics), item properties (IRT category response curves or item information curves, slope parameters, threshold values);

- **Standard errors** over the range of scores (e.g. in a figure);

If DIF testing has been performed (not a minimum requirement):

- Statistical method used (e.g. ordinal logistic regression), criterion used for DIF (e.g. McFadden's R^2 change of 2%),
- Names of the items that showed non-uniform and uniform DIF
- Figures, showing the relation of Theta and the total item score for all items and for the items with DIF are recommended.

2.2 Language- or country-specific reference scores in the general population

Establishing reference scores is an important step in the translation and cultural adaptation of a scale. Reference scores provide anchors to interpret an individual's or a group's score in relation to those of others. Reference scores are valuable in assessing the impact of disease. Reference scores also enable comparisons within and between countries, to compare scores from a specific patient sample in one country with those from a matched patient sample in another. And reference scores permit comparisons of the relative benefits of different treatments of various diseases between centers in a country or between countries.

PROMIS scores are always expressed in relation to the mean score of the general population. For most PROMIS instruments, a score of 50 represents the average for the United States general population. However, mean scores can differ among countries as a result of the translation or because of cultural differences. Therefore, **language- or country-specific reference scores of short forms and item banks are desired**.

Design: Cross-sectional study

Study population: for obtaining reference scores **the study population should be a representative sample from the general population**, with regard to age, gender, race, education, and geographic region. Inclusion criteria are: resident of the country of interest; age 8-17 for the pediatric item banks and 18+ for the adult item banks. Exclusion criteria are: insufficient command of the language of interest; insufficient cognitive ability to complete the items; no informed consent.

Different methods can be used to obtain a representative sample, such as random digit dialing, drawing samples from municipal population registries, or using an Internet polling panel. It is recommended to oversample subpopulations with lower expected response rates, such as ethnic minorities. Using census data (e.g. obtained from a national Bureau of Statistics) weights can be generated to compensate for non-response and non-coverage of the sample obtained [8]. Instead of drawing a random population sample, it may be more efficient to draw an equal number of people from relevant strata (i.e. strata based on age, gender, race, education, and geographic region). This will provide more power to calibrate the items in subgroups. Weights based on census data can then be used to estimate reference scores for the general population.

Sample size: In the literature, there is little guidance on required sample sizes for obtaining reference scores. Sample sizes of at least 120 subjects (after partitioning in relevant subgroups) have been recommended for obtaining references scores for laboratory values, to reduce the effect of extreme values [9]. The Dutch Institute for Psychologists recommends sample sizes of at least 300-

400 for tests used for individual decision-making [10]. For intelligence tests, often sample sizes of about 1000 are being used.

Data collection: All respondents should complete all items of a short form or item bank. If multiple item banks are administered, one may need multiple study populations who complete different item banks. PROMIS investigators recommend a maximum of 150 items be administered to a given respondent.

It is recommended to collect relevant characteristics of the study population (age, gender, race, and the presence of chronic diseases) for descriptive purposes and for calculating reference scores for relevant subgroups.

Analyses: References scores are described as mean (standard deviation (SD)) scores for the general population and for demographic subgroups. If necessary, scores should be weighted by census data.

2.3 Content validity of the item bank and cultural adaptation

Content validity refers to the relevance and comprehensiveness of the items included in an instrument. Items should be relevant for the construct, population and aim of the measurement application and no important items should be missing. Culture may influence perceptions of the meaning of constructs like fatigue or depression. Differences may especially exist between countries with large cultural differences (e.g. African or Asian countries as compared to Western countries). As a consequence, some items developed in the US may be less relevant in other countries or items important for a specific culture may be missing. Therefore, it is recommended to perform qualitative research to evaluate the relevance and comprehensiveness of the PROMIS items banks in new countries.

Cultural differences in the relevance of PRO items have been shown in the literature. For example, the PROMIS item “Does your health now limit you in putting a trash bag outside?” was considered irrelevant in the Netherlands because trash bags are hardly used anymore in the Netherlands [11]. Hoopman et al. found that some Muslim people had difficulty using the standard response categories of the question, ‘In general, how would you say your health is?’ Instead, they responded with ‘Hamdullilah’ (‘Thanks to God’), reflecting the widely held belief that one should accept one’s fate [12].

Important items that are relevant in other countries may not be included in the US PROMIS instruments. For example, cycling is a very important activity in the Netherlands but no item on cycling is included in the physical functioning item bank.

Finally, Some items may need to be country- or culturally adapted. For example, in China, school binders with rings are not commonly used, so the translation of the pediatric item “I could open the rings in school binders” was translated as “I could open binder clips”, which was considered to require a comparable level of dexterity and effort as opening the rings in binders, but is more relevant for Chinese-speaking children [13].

Most of these issues can be discovered and dealt with during the translation process. However, it may still be worthwhile to evaluate the relevance and comprehensiveness of the items of a short form or item bank in a separate study. Evaluating content validity is especially recommended in countries where large cultural differences with the US are expected.

Design: Qualitative research.

Population: The study population should be a relevant patient population for the item bank (e.g. patients with rheumatoid arthritis for the physical functioning item bank) or a sample from the general population. Experts in the field (e.g. health care providers, researchers) can be included to evaluate the relevance of the items for the construct to be measured.

Sample size: There are no power calculations or quantitative sample size estimations algorithms in qualitative research. Interviews should continue until ''saturation'' has been reached. This is the point whereby additional interviews are not expected to yield new or valuable information. It has been suggested in the literature that most projects reach saturation after conducting between 4 and 6 focus groups or 12 interviews [14].

Data collection: Participants should be asked about the relevance and comprehensiveness of all items and asked whether important items are missing. Focus groups or individual semi-structured interviews can be used. Ideally, all interviews should be conducted by the same facilitator to help maintain consistency in elicitation and evaluation techniques across interviews.

Analyses: qualitative analyses can be used if appropriate [14]. If necessary, new items should be developed and tested as described in the PROMIS standards [1].

Criteria for good content validity:
Evidence should be provided that patients and experts consider the content of the PROMIS instruments relevant and comprehensive for the construct, population, and aim of the measurement application.

2.4 Further psychometric evaluation of PROMIS instruments in specific patient populations

Validation is a continuous process and measurement properties are dependent upon the population in which an instrument is being used. Therefore, it is recommended to perform further psychometric analyses of PROMIS instruments after translation and accumulate evidence of validity over time. This concerns further calibration of the items, evaluation of **construct validity**, **internal consistency** (short forms only), **test-retest reliability** and **measurement error** of short forms and CATs. It is also recommended to test for **DIF among different patient groups**. It is recommended to use the PROMIS standards document [1] and the COSMIN standards for the design and analyses of studies on measurement properties [15].

Design: Cross-sectional study for item calibrations, evaluating construct validity and internal consistency and a test-retest design for evaluating reliability and measurement error. The test-retest period should be long enough to prevent patients from remembering their previous score, but short enough to assume that patients have not changed in the construct(s) being measured with the PROMIS instrument(s) under study. Usually a period of 1-2 weeks is recommended for Patient-Reported Outcome Measures (PROMs) [16, 17]. Patients should not be treated between the test and retest.

Population: The study population should be one or more relevant patient populations for the item bank (e.g. patients with mental diseases for the item banks anxiety and depression).

Sample size: For item calibrations, PROMIS recommends a minimum of 500 subjects per item (i.e. each item should have been completed by at least 500 subjects). For DIF analyses, a minimum of 200 subjects per group has been recommended [2]. For evaluating construct validity and test-retest reliability, COSMIN considers a sample size of 50 people as good and 100 as excellent [18]. If differences between subgroups are evaluated as evidence for construct validity, subgroups of at least 30 patients are recommended.

Data collection: All respondents should complete the PROMIS shorts forms or CATs of interest. It is recommended to collect relevant demographic variables (age, gender, relevant disease characteristics) for descriptive purposes and for evaluating relevant differences in scores between subgroups as evidence for construct validity.

For evaluating construct validity legacy instrument(s) should be included that measure the same construct as the item bank. It is recommend to include the most commonly used PROM or the PROM with the best measurement properties in the selected patient population (e.g. the WOMAC or KOOS subscale physical functioning for comparison with the PROMIS item bank physical functioning in patients with hip/knee osteoarthritis). It may also be useful to include the legacy instrument(s) in the retest, to enable direct comparison of reliability and measurement error of the PROMIS instrument versus the legacy instrument(s).

Analyses: Item calibrations and DIF analyses should be performed as described in the PROMIS standards document [1]. Internal consistency is only relevant for short forms and can be assessed by calculating Cronbach's alphas. **Construct validity should be assessed by testing predefined hypotheses** about expected correlations between PROMIS instruments and legacy instruments or about expected differences in PROMIS scores between relevant subgroups. **Test-retest reliability should be assessed by calculating the Intraclass Correlation Coefficient (ICC)** using a two-way random effects model for absolute agreement. **Measurement error should be assessed by calculating the Standard Error of Measurement (SEM)** as the square root of the variance between measurements and the error variance from the ICC or by calculating the Limits of Agreement [16].

To facilitate the interpretability of PROMIS scores it is recommended to present mean (SD) scores in relevant (sub)groups and to calculate the Smallest Detectable Change as 1.96*√2*SEM [16].

To help researchers take advantage of the measurement properties of the PROMIS instruments while maintaining continuity with previous research, crosswalk tables can be developed and tested to transform scores from legacy instruments to PROMIS scores. An example can be found in the study of Askew et al. [19].

Criteria for good measurement properties: PROMIS endorses **the minimum standards for patient-reported outcome measures** used in patient-centered outcomes and comparative effectiveness research that were recently **published by the International Society for Quality of Life Research (ISOQOL)** [20]. These were partly based on criteria which are often used in systematic reviews of PROMs [21]. Cronbach's alphas and ICCs should preferably be at or above 0.70. For construct validity correlations with instruments measuring similar constructs should be higher than correlations with instruments measuring dissimilar constructs. It has been recommended that 75% of the results should support the predefined hypotheses.

2.5 Responsiveness and Minimal Important Change of PROMIS instruments in relevant patient populations

Responsiveness is defined as the ability of an instrument to detect change over time in the construct to be measured [22]. Responsiveness is an important measurement property for PROMIS instruments that are being used repeatedly over time to measure changes in patient-reported health status. It has been shown that PROMIS instruments are more responsive than commonly used PROMs [23]. However, like other measurement properties, responsiveness is dependent upon the population in which an instrument is being used. Therefore, it is recommended to evaluate the responsiveness of PROMIS instruments after translation in one or more relevant patient populations. The same study can be used to assess the Minimal Important Change (MIC) of PROMIS instruments. The MIC is the smallest change in score that patients consider to be important. Knowing the MIC facilitates the interpretability of PROMIS change scores and can also be used to conduct responder analyses in clinical trials [24].

Design: Longitudinal study with at least two measurements. **At least part of the patients should experience change** in the construct being measured with the PROMIS instrument. If the sample exists of patients with chronic progressive diseases, or if an intervention is given, it is likely that at least part of the patients have changed. Also a global rating scale of change can be used to ask patients if they consider themselves as changed.

Population: The study population should be a relevant population for the item bank (e.g. patients with rheumatoid arthritis for the physical functioning item bank).

Sample size: COSMIN considers a sample size of 50 people as good and 100 as excellent [18]. If differences between changes in subgroups are evaluated as evidence for responsiveness, subgroups of at least 30 patients are recommended.

Data collection: All respondents should complete the PROMIS shorts forms or CATs of interest at all time points. It is recommended to collect relevant demographic variables (age, gender) at baseline for descriptive purposes, and relevant disease characteristics at all time points for descriptive purposes and for evaluating differences in change scores on the PROMIS instruments between relevant subgroups.

It is recommended to include legacy instrument(s) at all time points that measure the same construct as the PROMIS instruments so that changes in scores on the PROMIS instruments can be compared with changes in scores on the legacy instrument(s). It is recommend to include the most commonly used PROM or the PROM with the best measurement properties in the selected patient population (e.g. the Health Assessment Questionnaire (HAQ) for comparison with the PROMIS item bank physical functioning in patients with rheumatoid arthritis).

It is also recommended to include a global rating scale of change (also called anchor or external criterion) at the follow-up measurement(s) to ask patients if they consider themselves as changed (e.g. on a 5-point scale). The anchor should ask about change in the same construct that the PROMIS instrument measures (e.g. ask about change in physical functioning for comparison with changes in the PROMIS item bank physical functioning). If multiple PROMIS instruments are being evaluated, multiple anchors should be used.

Analyses: Responsiveness should be assessed by **testing predefined hypotheses about expected correlations between changes** in PROMIS instruments and changes in legacy instruments, between changes in PROMIS instruments and the anchor questions, or about expected differences in PROMIS change scores between relevant subgroups [25]. In addition, the ROC method can be used to evaluated how well PROMIS instruments can discriminate between patients who have changed and patients who have not changed based on the anchors, legacy instruments, or changes in relevant disease characteristics [26].

For assessing the MIC anchor-based methods are recommended in which change scores on the PROMIS instruments are being compared to one or more anchors [27]. De Vet et al. recommend to plot the distributions of the change scores in patients who have been importantly improved and patients who have not been changed to see how well the instrument distinguishes between these groups [28].

References

1. (2013). *PROMIS® Instrument Development and Validation Scientific Standards. www.nihpromis.org/science/publications.*

2. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de, G. A., Groenvold, M. et al . (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *J Clin Epidemiol, 62*, 288-295.

3. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *J Stat Softw, 39*, 1-30.

4. Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2012). Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Qual Life Res,*

5. Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How Big Is Big Enough? Sample Size Requirements for CAST Item Parameter Estimation. *Applied Measurement in Education, 19*, 241-255.

6. Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

7. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S. et al . (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol, 63*, 1179-1194.

8. Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W. et al . (2010). Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *J Clin Epidemiol, 63*, 1169-1178.

9. Gräsbeck, R. & Saris, N. E. (1969). Establishment and use of normal values. *Scand J Clin Lab Invest, 26*, 62-63.

10. Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010) COTAN beoordelingssysteem voor de kwaliteit van tests.

11. Terwee, C. B., Roorda, L. D., de Vet, H. C. W., Dekker, J., Westhovens, R., van Leeuwen, J. et al . (2013). Dutch-Flemish translation of 17 item banks from the Patient Reported Outcomes Measurement Information System (PROMIS). *Submitted for publication.*

12. Hoopman, R., Terwee, C. B., Muller, M. J., Ory, F. G., & Aaronson, N. K. (2009). Methodological challenges in quality of life research among Turkish and Moroccan ethnic minority cancer patients: translation, recruitment and ethical issues. *Ethn Health, 14*, 237-253.

13. Liu, Y., Hinds, P. S., Wang, J., Correia, H., Du, S., Ding, J. et al . (2013). Translation and linguistic validation of the Pediatric Patient-Reported Outcomes Measurement Information System measures into simplified Chinese using cognitive interviewing methodology. *Cancer Nurs, 36*, 368-376.

14. Brod, M., Tesler, L. A., & Christensen, T. L. (2009). Qualitative research and content validity: developing best practices based on science and experience. *Quality of Life Research, 18*, 1263-1278.

15. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L. et al . (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research, 19*, 539-549.

16. de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine*. (Cambridge: Cambridge University Press)

17. Streiner, D. L. & Norman, G. R. (2003). *Health measurement scales. A practical guide to their development and use*. (New York: Oxford University Press)

18. Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res, 21*, 651-657.

19. Askew, R. L., Kim, J., Chung, H., Cook, K. F., Johnson, K. L., & Amtmann, D. (2013). Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. *Qual Life Res,*

20. Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F. et al . (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res*, *8*, 1889-905.

21. Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J. et al . (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol, 60*, 34-42.

22. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L. et al . (2010). International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. *Journal of Clinical Epidemiology, 63*, 737-745.

23. Fries, J. F., Krishnan, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Res Ther, 13*, R147.

24. Schunemann, H. J., Akl, E. A., & Guyatt, G. H. (2006). Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health Qual Life Outcomes, 4*, 62.

25. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L. et al . (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodeology, 10*, 22.

26. Deyo, R. A. & Centor, R. M. (1986). Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. *Journal of Chronic Diseases, 39*, 897-906.

27. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol, 61*, 102-109.

28. de Vet, H. C., Ostelo, R. W., Terwee, C. B., van der, R. N., Knol, D. L., Beckerman, H. et al . (2007). Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res, 16*, 131-142.