Published in final edited form as:
Qual Life Res. 2010 May ; 19(4): 595–607. doi:10.1007/s11136-010-9619-3.

An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales

Debra E. Irwin,
Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Brian Stucky,
Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Michelle M. Langer,
National Board of Medical Examiners, Philadelphia, PA, USA

David Thissen,
Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Esi Morgan DeWitt,
Department of Pediatrics, Duke University Medical Center, Durham, NC, USA

Jin-Shei Lai,
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

James W. Varni,
Department of Pediatrics, College of Medicine, Department of Landscape Architecture and Urban Planning, College of Architecture, Texas A&M University, College Station, TX, USA

Karin Yeatts, and
Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Darren A. DeWalt
Division of General Medicine and Clinical Epidemiology, Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Debra E. Irwin: dirwin@email.unc.edu

Abstract

Purpose—The Patient-Reported Outcomes Measurement Information System (PROMIS) aims to develop self-reported item banks for clinical research. The PROMIS pediatrics (aged 8–17) project focuses on the development of item banks across several health domains (physical function, pain, fatigue, emotional distress, social role relationships, and asthma symptoms). The psychometric properties of the anxiety and depressive symptom item banks are described.

Methods—Participants (n = 1,529) were recruited in public school settings, hospital-based outpatient and subspecialty pediatrics clinics. The anxiety (k = 18) and depressive symptoms (k = 21) items were split between two test administration forms. Hierarchical confirmatory factor-analytic models (CFA) were conducted to evaluate scale dimensionality and local dependence. IRT analyses were then used to finalize item banks and short forms.

Correspondence to: Debra E. Irwin, dirwin@email.unc.edu.

Results—CFA results confirmed that anxiety and depressive symptoms are separate constructs and indicative of negative affect. Items with local dependence and DIF were removed resulting in 15 anxiety and 14 depressive symptoms items. The psychometric differences between short forms and simulated computer adaptive tests are presented.

Conclusions—PROMIS pediatric item banks were developed to provide efficient assessment of health-related quality of life domains. This sample provides initial calibrations of anxiety and depressive symptoms item banks and creates PROMIS pediatric instruments, version 1.0.

Keywords

PROMIS; Anxiety; Depressive symptoms; HRQOL; PRO; Scale development; Surveys; Pediatrics

Background

The Patient-Reported Outcomes Measurement Information System (PROMIS) project, a National Institute of Health Roadmap for Medical Research initiative, was developed to advance the science and application of patient-reported outcomes (PRO) in chronic diseases [1]. One main goal of the PROMIS initiative is to develop a set of PRO item banks and computerized adaptive tests for the clinical research community. The PROMIS Pediatric project focused on the development of self-report PRO item banks across several health domains for youth aged 8–17. The generic health domains are important across a variety of illnesses, and include physical function, pain, fatigue, emotional distress, and social function [2]. Additionally, one disease-specific item bank was developed for children with asthma to measure disease-related symptoms.

Emotional distress commonly refers to unpleasant feelings or emotions that are experienced privately and, therefore, are good candidates for assessment as PROs. Emotional distress among children is partially comprised of feelings of anxiety, depression, and anger [3]. The emotional distress domains of anxiety and depressive symptoms are the focus of this manuscript.

Symptoms that best differentiate anxiety are those that reflect autonomic arousal and the experience of threat. Children often experience these feelings in a variety of contexts specific to their environment of home, school, and social activities [3]. The PROMIS pediatric item bank for anxiety focuses on fear (e.g., fearfulness), anxious misery (e.g., worry), and hyperarousal (e.g., nervousness).

Depressive symptoms among children often include feelings of hopelessness, helplessness, and worthlessness [3]. The PROMIS pediatric item bank for depressive symptoms focuses on negative mood (e.g., sadness), anhedonia (e.g., loss of interest), negative views of the self (e.g., worthlessness, low self-esteem), and negative social cognition (e.g., loneliness, interpersonal alienation). This item bank is best characterized as depressive symptoms rather than as a complete diagnostic test for depression.

PROMIS pediatric item banks were developed using a strategic item generation methodology adopted by the PROMIS Network [4]. Six phases of item development were implemented: *identification of existing items*, *item classification and selection*, *item review and revision*, *focus group input on domain coverage*, *cognitive interviews with individual items*, and *final revision before field testing*. Identification of items refers to the systematic search for existing items in currently available pediatric scales [2, 4–6]. Items successfully screened through the process were sent to field testing. The final PROMIS pediatric item set contained 15 anxiety and 14 depressive symptom items.

A limited number of generic self-report health-related quality of life (HRQOL) instruments exist for use in pediatric populations, and most attempt to measure at least some aspect of emotional distress [7]. The vast majority of these have utilized classical test theory and few have taken advantage of item response theory (IRT) analysis in the scale development process [8]. PROMIS psychometric analyses focused on determining scale dimensionality and detecting sources of local dependence and also considered final item selection using IRT analyses. The primary objective of this paper is to describe the IRT analyses of the PROMIS pediatric anxiety and depressive symptoms item banks and the measurement properties of the new PROMIS pediatric anxiety and depressive symptoms scales that resulted from these IRT analyses, including investigations of scale dimensionality, sources of local dependence, and differential item functioning (DIF).

Methods

Sampling plan

Participants from central North Carolina and Texas were recruited in hospital-based outpatient general pediatrics and subspecialty clinics and in public school settings between January 2007 and May 2008. School-based participants were recruited through elementary after school programs as well as middle and high school required health classes. Parental informed consent and minor assent were obtained for all children taking the survey. A more detailed description of the study design is provided elsewhere [9].

The PROMIS anxiety and depressive symptoms items were randomly split between two test administration forms (Form 1 contained 9 anxiety items and 10 depressive symptom items; Form 2 contained 9 anxiety items and 11 depressive symptom items). Children were randomly assigned to complete one of the testing forms. Each of the anxiety and depressive symptoms PROMIS pediatric items was administered to at least 759 respondents. This sampling plan was developed for collecting responses to candidate items from the targeted PROMIS domains and accommodated multiple objectives including: (1) confirm the factor structure of the domains; (2) evaluate items for local dependence (LD) and DIF; and (3) calibrate the items for each domain using IRT.

All of these emotional distress items had a 7-day recall period and used standardized 5-point response options (never, almost never, sometimes, often, almost always). Table 1 shows the anxiety and depressive symptoms items administered during the testing.

Statistical and psychometric methods

Data analysis followed the sequence of procedures presented by Reeve et al. [10] in their description of plans for psychometric evaluation and calibration of health-related quality of life item banks for PROMIS. First, traditional descriptive statistics were computed, as a check on data entry and validity and to verify that there were no empty (zero frequency) response categories for any item. These statistics included the frequencies and proportions in each item response category and the correlation of the item scores with the total summed score.

Second, as a check on the assumptions of the unidimensional IRT model to be used, the dimensionality of individual differences on the anxiety and depressive symptoms item sets was examined using confirmatory factor analysis (CFA; e.g., bifactor analysis) of the inter-item polychoric correlation matrices. These analyses were performed using the “weighted least squares with robust standard errors, mean- and variance-adjusted” (WLSMV) algorithm [11] as implemented in the software M*plus* [12]. Fitting additional factors, over and above those indicated by the design of the questionnaire, and residual correlations significantly greater than zero, served as indices of local dependence (LD) for pairs or small

numbers of items that violate the local independence assumption of unidimensional IRT [13]. If a pair of items exhibited LD, one item from the pair was set aside.

Third, within the sets of items for which unidimensionality had been confirmed using CFA, the items were "calibrated" by fitting Samejima's Graded Response Model [14, 15] using the software Multilog [16]. This model characterizes each item with a *slope* or *discrimination* parameter (a), that reflects the degree of association of the item responses with the latent construct being measured, and four threshold parameters (b_k) (for five-alternative items), that indicate the level of anxiety or depressive symptoms at which a response in a particular category or higher becomes likely. This model has been selected for the PROMIS scales [10]. The goodness of fit of the IRT model to the data was examined using Orlando and Thissen's [17, 18] *SS* X^2 statistic as generalized by Bjorner et al. [19] for polytomous response data. Because *SS* X^2 is a goodness-of-fit statistic, a non-significant value is the desirable outcome, indicating adequate fit of the model to the data.

Fourth, the possibility of differential item functioning (DIF) was investigated for each item on each scale using the IRT-LR DIF detection procedure [20] as implemented in the software IRTLRDIF [21]. DIF indicates that the relation of the item responses with the latent variable being measured differs between two (most often demographic) groups. Such a difference implies that some other factor, related to group membership but different from the construct being measured, had an influence on the item responses, violating the IRT assumption of unidimensionality. In the present data, the only demographic background variable that divides the sample into two groups that are sufficiently large to examine DIF is gender, so the DIF analysis was done separating the data into responses from boys and girls. IRT-LR DIF detection provides a X^2-distributed test statistic; again, a non-significant value is the desirable outcome, indicating a lack of detectable DIF. We used the Benjamini–Hochberg [22, 23] procedure to control for the multiplicity of comparisons involved in checking each item for DIF using $\alpha = 0.05$, and graphical methods, as suggested by Steinberg and Thissen [24] to evaluate effect size when DIF was detected.

Item pools were generated by combining the remaining items across the form. The linking procedure used, called "common population linking" or "randomly equivalent groups," is based on calibrating multiple test forms from a common population and is widely used in educational testing [25, 26]. This technique is appropriate in this situation because items were randomly assigned to test forms, and test forms were randomly assigned to individuals. These procedures enabled the research team to administer nearly 300 items across multiple forms and domains without exceeding 70 items on any particular form.

Fifth, after the final item pools were selected, confirmatory factor analysis (CFA) of the inter-item polychoric correlation matrix among the remaining, selected items was used to ensure that the latent variables underlying the item responses for the anxiety and depressive symptoms were unidimensional in the final item pools. These analyses were performed using the DWLS algorithm as implemented in the software LISREL [27].

Finally, IRT scores for the scales are based on the GRM parameters after the scales are assembled [28]. All IRT-based scores are relative to some reference group [29]; in this case, the reference group is the subset of the sample from the NC. While IRT scale scores may be based either on item response patterns or summed scores, we expect most often scale scores based on summed scores will be used; score translation tables for that purpose are provided in the "Appendix" and Table 7.

Results

Test forms containing anxiety and depressive symptoms PROMIS pediatric items were completed by a total of 1,529 respondents. The sample was about 52% female and 58% were children aged 8–12. Fifty-nine percent were Caucasian, 21% black, 6% multi-racial, and 14% other races (Asian/Pacific Islanders, Native Americans, and Other Races). Seventeen percent of the sample was of Hispanic ethnicity. The vast majority of the adults providing informed consent for the children were parents of the child (92%) or grandparents (4%). The educational attainment of these parents or guardians ranged from less than high school (7%) to advanced degree (15%) with 26% reporting a college degree, 33% some college, and 20% a high school diploma. Approximately 23% of the children participating in the survey had a chronic illness diagnosis during the past 6 months (Table 2).

Using CFA to examine the dimensionality of the anxiety and depressive symptoms item sets on the two-item tryout forms involved fitting a number of models; the factor loadings for representative models that fit the data reasonably well are shown in Tables 3 and 4. The items in both tables are sorted to group together items with similar statistical properties. Both tables show modified bi-factor models, comprising a general factor with loadings for all items, two group-specific factors with non-zero loadings for either the anxiety or the depressive items (that much is a bi-factor model), plus a smaller group factor (in Table 3) and residual correlations. The latter components augment the bi-factor model and represent local dependence between pairs, or a triplet, of items. Indicators of goodness of fit suggest both models fit the data, using as standards suggested by Reeve et al. [10]: For the model in Table 3, $X^2(76) = 248$, CFI = 0.95, TLI = 0.99, RMSEA = 0.06; for the model in Table 4, $X^2(91) = 194$, CFI = 0.97, TLI = 0.99, RMSEA = 0.04.

These models answer two questions that arise in the context of scale construction using these items. The first question answered, “Is negative affect unidimensional, or is there distinguishable individual difference variation corresponding to the anxiety items distinct from the depressive symptoms items?” The fact that there are substantial loadings that differ significantly from zero on the group-specific factor for the depressive symptoms items in Table 3, and on that for the anxiety items in Table 4, indicates that the covariation among the item responses cannot be adequately explained with a theory that there is a single negative-affect dimension of individual differences underlying responses to all of the items. Both anxiety and depressive symptoms have their own unique components. (It is a curious fact that the general factor in Table 3 is anxiety-dominated negative affect, leaving little unique for the anxiety group-specific factor; and the general factor in Table 4 is dominated by depressive symptoms, leaving little unique for the depressive symptoms group-specific factor. However, that is simply an illustration of the fact that the composition of latent variables in factor analysis is highly dependent on the properties of the item set being analyzed.)

A second question answered by these factor-analytic results is: “Are the items conditionally independent, given the combination of the general factor and group-specific factors for anxiety and depressive symptoms?” The answer is: “Not all of them.” In Table 3, there is a cluster of items that involve being “scared” or “afraid” with responses that are more correlated than is expected given the general factor and the anxiety-specific factor, and there are four more pairs of items with significant residual correlations. In Table 4, there are two more locally dependent pairs. Items in these pairs or triplets are (in part) like “asking the same question twice” (a common sense description of LD), so we will include only one item from the triplet, or from each pair, in each final item pool so that each pool comprises locally independent items.

The model in Table 4 includes a residual correlation between one of the anxiety items, “I felt worried,” and one of the depressive symptoms items, “I felt stressed.” Because those two items are ultimately destined for the distinct anxiety and depressive symptoms item pools, that residual correlation does not constitute a violation of the assumption of local independence within either scale, so that residual correlation is ignored.

We used the results of additional CFA analyses, not shown here, to examine the possibility that the factor structure for the anxiety and depressive symptoms might be different for younger children than it is for adolescents. We divided the sample approximately in half by age and estimated parameters for models like those shown in Tables 3 and 4, and found that there were no substantial differences between the parameter estimates for younger (aged 8–12) versus older (aged 13–17) children (data not shown).

Two items that were included in the anxiety sets for the CFA, “I felt afraid or scared” and “I worry about what will happen to me” are actually legacy items (from the Pediatric Quality of Life Inventory™ (PedsQL™) Version 4.0 Generic Core Scales; [30]). Those items were included in the CFA for reference purposes and they were set aside before item calibration for the PROMIS scales. Then we calibrated the remaining items on forms 1 and 2 separately for the anxiety and depressive symptoms dimensions. To avoid deleterious effects of locally dependent item pairs on the item parameter estimates [31], we computed the item parameter estimates twice for each form, including only one member of each LD pair in the item set at a time. That produced two sets of item parameters for the non-locally dependent items; to reduce capitalization on chance, we selected the set with the lower slope (*a*) parameter from those pairs. The values of the item parameter estimates and the *SS* X^2 item fit statistics are shown in Tables 5 and 6. In those tables, the items that remain in the final item pools are sorted in order of decreasing discrimination (*a*), so the generally best indicators of anxiety and depressive symptoms are near the top of the tables.

For the anxiety items, after using the Benjamini–Hochberg correction for multiplicity, none of the items exhibited significant lack of fit as indicated by the *SS* X^2 statistic. Similarly, none of the gender DIF tests (also shown in Table 5) was significant when adjusted for multiplicity. So we set aside the three items listed near the bottom of Table 5; two of those were the less-discriminating items in pairs the CFA had indicated exhibited LD. The third item, “I felt relaxed,” was set aside because it was not nearly as discriminating as the other items, and, upon reflection, even reverse-scored as it is, is probably not a particularly specific indicator of anxiety. That left the 15-item set in the upper part of Table 5 as the pediatric anxiety item pool.

Similarly, for the depressive items, after using the Benjamini–Hochberg correction for multiplicity, none of the items exhibited significant lack of fit as indicated by the *SS* X^2 statistic. However, three items listed near the bottom of Table 6 exhibited significant DIF. Unsurprisingly, two of those items involved “crying,” which usually exhibits DIF between genders on depression scales [32–34]. The reason for DIF for the item “I felt so bad that I didn’t want to do anything” is perhaps not so clear; however, the item was set aside due to the magnitude of the effect size when depicted graphically. As shown in Table 6, three additional depressive symptoms items were set aside because they were the less-discriminating members of three LD pairs (as detected with residual correlations in the CFA). Finally, the item “I was bored” was set aside because it appeared to be a poor indicator of depressive symptoms. The remaining 14 items in the upper half of Table 6 are the pediatric PRO-MIS depressive symptoms item pool.

To ensure that the final item pools for anxiety and depressive symptoms were unidimensional, and to estimate the correlation between those two latent variables, two-

factor simple-structure CFA models were fitted to the selected anxiety and depressive symptoms on Forms 1 and 2. In each of these models, all of the selected anxiety items loaded on the anxiety factor, with zero loadings on the depressive symptoms factor, and all of the selected depressive symptoms items loaded on the depressive symptoms factor, with zero loadings on the anxiety factor. The goodness of fit of these models was satisfactory for both forms (for Form 1: $X^2(53) = 247$, CFI = 0.98, TLI = 0.97, RMSEA = 0.07; for Form 2: $X^2(118) = 289$, CFI = 0.99, TLI = 0.99, RMSEA = 0.04). The estimated correlations between the latent variables depressive symptoms and anxiety were 0.82 (SE = 0.02) for Form 1, and 0.85 (SE = 0.02) for Form 2. Additional separate one-factor analyses of the selected items on each scale for each form for boys and girls fit well, with no indication of further local dependence.

The upper panel of Fig. 1 shows the test information function for the anxiety item pool, and for the eight-item subset of the pool that is most informative near the middle of the distribution, and the lower panel of Fig. 1 shows the corresponding curves for the depressive symptoms item pool. (The *x*-axes in Fig. 1 are labeled using the *T*-score scale on which scores from PROMIS scales are reported, with a mean of 50 and a standard deviation of 10 for the reference population.) Test information is the inverse of the variance of measurement (the squared standard error of measurement), so a value of 6.67 for information corresponds with a standard error of measurement of approximately 0.4 standard units (for standardized test scores, or 4 for *T*-scores), and that in turn corresponds to a reliability coefficient of approximately 0.85. For anxiety and depressive symptoms, eight items are sufficient to provide measurement with that precision for the range of scores from the low 40s to nearly 80 on the *T*-score scale. Thus, for many purposes, we recommend using the short eight-item forms that are listed in the Appendix, with the corresponding summed score to scale score translation tables based on the IRT model.

For situations that require more precision of measurement, the complete item pools are in Tables 5 and 6, with the item parameters that can be used to compute IRT response pattern scores or the scale scores for summed scores for any other (larger or smaller) sets of the items, all on a comparable scale. In addition, the item pools are available from the Assessment Center at www.nihpromis.org.

To consider whether adaptive testing might be useful, we computed the test information curves for the most informative set of items from the anxiety and depressive symptoms item pools at *T*-scores of 30, 40, 50, 60, and 70. These curves are shown in the upper and lower panels, respectively, of Fig. 2. These curves basically answer the questions: “How much more information can be gained by choosing a different set of ‘best items’ for different score levels (adapting), given this item pool?” and “To what extent are different sets of items ‘best’ at different levels of anxiety or depressive symptoms?” The answer to both questions, as shown in Fig. 2, is “not very much.” The items measure anxiety and depressive symptoms well between T-scores in the low 40s and 80. Within that range, to a large extent the same items are most informative, and outside that range adapting, even to the extent of using all of the questions in the pool, adds little precision. Thus, we have not further evaluated the idea of using adaptive testing with these item pools; however, the pools and item parameters are available, and the PROMIS Assessment Center software has the capability of administering these items as a CAT if that is desired.

Discussion

This study led to the development of new anxiety and depressive symptoms item banks for use in measuring pediatric PROs. After determining scale dimensionality, items with local dependence and DIF were next identified and removed resulting in final item banks with 15

anxiety items and 14 depressive symptom items, allowing a variety of possible approaches to scoring that can be tailored to meet the goals of the end user.

Several generic self-report HRQOL instruments exist for use in pediatric populations and most attempt to measure at least some aspect of emotional distress. The vast majority of the generic pediatric HRQOL measures for anxiety and depressive symptoms utilized classical test theory and few have taken advantage of IRT analysis in the scale development process [8]. PROMIS psychometric analyses focused on determining the scale dimensionality and detecting sources of local dependence and considered final item selection using IRT analyses. Like PROMIS, two of these newer instruments, KIDSCREEN and PedsQL, utilized qualitative research methods for incorporating the child's perspective during the development process [35, 36].

One major challenge prior to applying IRT models to the measurement of emotional distress is resolving issues of dimensionality. Conventional wisdom is that emotional distress scales are less likely to fit unidimensional models [37]. Often items are sampled from multiple domains (e.g., mood, behavior, somatic symptoms) in order to capture a comprehensive set of latent construct indications. Hence, it is common to observe higher correlations within domains than is expected under the conditional independence assumption of unidimensional IRT models [38]. One of the initial steps for this project was to develop multidimensional conceptual frameworks that were informed by previous empirical (e.g., factor analytic) and theoretical work as well as to determine the level of resolution at which unidimensional scales could be derived from the domains [2, 4–6]. Three constructs of emotional distress were conceptualized: depressive symptoms, anxiety, and anger. These results of unidimensionality are consistent with a recent meta-analysis [39] and other published studies [40–45].

It remains a question for comparative validity studies to determine which of these scales might be most valid for any particular use. Both the KIDSCREEN *Moods and Emotions* scale (7 items) and the PedsQL *Emotional Functioning* scale (5 items) are shorter than the PROMIS depressive symptoms or anxiety 8-item short forms and provided less reliable measurement in this item calibration sample: Coefficient alpha for the KIDSCREEN *Moods and Emotions* scale in this sample was 0.83, and for the PedsQL *Emotional Functioning* scale, it was 0.74. While the PROMIS scales provide separate scores for depressive symptoms and anxiety, the PedsQL *Emotional Functioning* scale includes items that indicate depression, anxiety, and anger while the KIDSCREEN *Moods and Emotions* scale largely measures depressive symptoms, with one item that may indicate anxiety. It also remains a question for future validity studies to determine the usefulness of separate scores for depressive symptoms and anxiety: These two constructs are highly correlated; however, it may be that either one or both are responsive to any particular treatment or that they are affected separately or together by any particular condition. The separate scores of the PROMIS pediatric measures permit study of those questions.

Utilizing IRT analysis to identify final items ultimately offers more flexibility for future users of the item banks. This approach allows researchers the opportunity to select the most useful items for their study design. We proposed 8-item short forms; however, a smaller subset of items from the item bank can also be used and scored on the same metric as the larger set.

By administering the items spread over several test forms, we are unable to perform factor analyses across each entire bank. This limitation makes it impossible to ensure that items from different forms do not exhibit local dependence. Additionally, it is possible that factor analyses for each domain would turn out differently if the items were analyzed all together.

Instead, factor analysis was conducted over the subgroups of items tested on each form. Because the items were created to fill content from qualitative work and then were randomly allocated to each test form, the different test forms can be viewed as replications. By having replicated factor analyses, our impressions of multidimensionality, when repeated across forms, increased our confidence in the factor-analytic results. We are currently performing cross-sectional testing using the entire item pools to verify these results.

The PROMIS pediatric item banks were developed to provide accurate and efficient assessment of important domains of HRQOL for children including anxiety and depressive symptoms. This sample provides initial calibrations of the PROMIS pediatric anxiety and depressive symptoms item banks and the creation of the corresponding PROMIS Pediatric instruments, version 1.0.

Abbreviations

PROMIS	Patient-reported outcomes measurement information system
PedsQL™	Pediatric quality of life inventory™
HRQOL	Health-related quality of life
PRO	Patient-reported outcomes
CFA	Confirmatory factor analysis
IRT	Item response theory
LD	Local dependence
DIF	Differential item function

References

1. Ader DN. Developing the patient-reported outcomes measurement information system (PROMIS). Medical Care. 2007; 45(Suppl 1):S1–S2. [PubMed: 18027399]
2. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Medical Care. 2007; 45(Suppl 1):S3–S11. [PubMed: 17443116]
3. [last accessed October 5, 2009.] http://www.nimh.nih.gov/health/topics/child-and-adolescent-mental-health/index.shtml
4. DeWalt D, Rothrock N, Yount S, Stone AA. PROMIS cooperative group: Evaluation of item candidates: The PROMIS qualitative item review. Medical Care. 2007; 45(Suppl 1):S12–S21. [PubMed: 17443114]
5. Walsh TR, Irwin DE, Meier A, Varni JW, DeWalt D. The use of focus groups in the development of the PROMIS pediatric item bank. Quality of Life Research. 2008; 17:725–735. [PubMed: 18427951]
6. Irwin DE, Varni JW, Yeatts K, DeWalt D. Cognitive interviewing methodology in the development of a pediatric item bank: A patient reported outcomes measurement information system (PROMIS) study. Health and Quality of Life Outcomes. 2009; 7(3):1–10. [PubMed: 19134191]
7. Matza LS, Swensen AR, Flood EM, Secnik K, Leidy NK. Assessment of health-related quality of life in children: A review of conceptual, methodological, and regulatory issues. Value in Health. 2004; 7:79–92. [PubMed: 14720133]
8. Ravens-Sieberer U, Erhart M, Wille N, Wetzel R, Nickel J, Bullinger M. Generic health-related quality-of-life assessment in children and adolescents: Methodological considerations. Pharmacoeconomics. 2006; 24(12):1199–1220. [PubMed: 17129075]

9. Irwin DE, Stucky BD, Thissen D, Morgan DeWitt E, Lai JS, Yeatts K, Varni JW, DeWalt DA. Sampling plan and patient characteristics of the PROMIS pediatrics large scale survey. Quality of Life Research manuscript under review. 2010
10. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). Medical Care. 2007; 45:S22–S31. [PubMed: 17443115]
11. Muthén, B.; du Toit, SHC.; Spisic, D. Robust inference using weighted least squared and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Los Angeles, CA: Muthén & Muthén; 1997. Unpublished technical report
12. Muthen, LK.; Muthen, BO. Mplus user's guide. 2. Los Angeles, CA: Muthen & Muthen; 2004.
13. Hill CD, Edwards MC, Thissen D, Langer MM, Wirth RJ, Burwinkle TM, et al. Practical issues in the application of item response theory: A demonstration using items from the pediatric quality of life inventory™ (PedsQL™) 4.0 generic core scales. Medical Care. 2007; 45:S39–S47. [PubMed: 17443118]
14. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph. No. 17. 1969
15. Samejima, F. Graded response model. In: van der Linden, WJ.; Hambleton, RK., editors. Handbook of modern item response theory. New York: Springer; 1997. p. 85-100.
16. du Toit, M., editor. IRT from SSI. Lincolnwood, IL: Scientific Software International; 2003.
17. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24:50–64.
18. Orlando M, Thissen D. Further examination of the performance of S-X^2, an item fit index for dichotomous item response theory models. Applied Psychological Measurement. 2003; 27:289–298.
19. Bjorner, JB.; Smith, KJ.; Edelen, MO.; Stone, C.; Thissen, D.; Sun, X. IRTFIT: A macro for item fit and local dependence tests under IRT models. Lincoln, RI: QualityMetric Incorporated; 2007.
20. Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. p. 67-113.
21. Thissen, D. IRTLRDIF: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory; The University of North Carolina; Chapel Hill: 2001.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 1995; 57:289–300.
23. Williams VSL, Jones LV, Tukey JW. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. Journal of Educational and Behavioral Statistics. 1999; 24:42–69.
24. 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. [PubMed: 17154754]
25. Kolen, MJ.; Brennan, RL. Test equating, scaling, and linking. 2. New York, NY: Springer; 2004.
26. Dorans NJ. Linking scores from multiple health outcome instruments. Quality of Life Research. 2007; 16(s1):85–94. [PubMed: 17286198]
27. Joreskog, KG.; Sorbom, D. LISREL 8.5. Lincolwood, IL: Scientific Software International, Inc; 2003.
28. Thissen, D.; Nelson, L.; Rosa, K.; McLeod, LD. Item response theory for items scored in more than two categories. In: Thissen, D.; Wainer, H., editors. Test scoring. Mahwah, NJ: Lawrence Erlbaum Associates; 2001. p. 141-186.
29. Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. Quality of Life Research. 2007; 16:109–116. [PubMed: 17294284]
30. Varni JW, Seid M, Kurtin PS. The PedsQL™ 4.0: Reliability and validity of the pediatric quality of life inventory™ version 4.0 generic core scales in healthy and patient populations. Medical Care. 2001; 39:800–812. [PubMed: 11468499]

31. Chen WH, Thissen D. Local dependence indices for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22:265–289.
32. Reeve, BB. Unpublished doctoral dissertation. University of North Carolina; Chapel Hill: 2000. Item- and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory.
33. Santor DA, Ramsay JO, Zuroff DC. Nonparametric item analyses of the beck depression inventory: Evaluating gender item bias and response option weights. Psychological Assessment. 1994; 6:255–270.
34. Schaeffer, NC. An application of item response theory to the measurement of depression. In: Clogg, CC., editor. Sociological methodology. Vol. 18. Washington, DC: American Sociological Association; 1998. p. 271-307.
35. Varni JW, Seid M, Rode CA. The PedsQL™: Measurement model for the pediatric quality of life inventory™. Medical Care. 1999; 37:126–139. [PubMed: 10024117]
36. Ravens-Sieberer U, Gosch A, Rajmil L, Erhart M, Bruil J, Duer W, et al. KIDSCREEN-52 quality of life measure for children and adolescents. Expert Review of Pharmacoeconomics and Outcomes Research. 2005; 5:353–364. [PubMed: 19807604]
37. Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, Grochocinski VJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. Psychiatric Services. 2008; 59:361–368. [PubMed: 18378832]
38. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Quality of Life Research. 2007; 16:95–108. [PubMed: 17530450]
39. Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. Journal of Clinical Psychology. 2006; 62:123–146. [PubMed: 16287149]
40. Watson D, Clark LA, Weber K, Assenheimer JA, Strauss ME, McCormick RA. Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptoms. Journal of Abnormal Psychology. 1995; 104:3–14. [PubMed: 7897050]
41. Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. Journal of Abnormal Psychology. 1995; 104:15–25. [PubMed: 7897037]
42. Clark LA, Watson D. Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. Journal of Abnormal Psychology. 1991; 100:316–336. [PubMed: 1918611]
43. Chorpita BF, Albano AM, Barlow DH. The structure of negative emotions in a clinical sample of children and adolescents. Journal of Abnormal Psychology. 1998; 107:74–85. [PubMed: 9505040]
44. Brown TA, Chorpita BF, Barlow DH. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. Journal of Abnormal Psychology. 1998; 107:179–192. [PubMed: 9604548]
45. Pilkonis PA, Reise SP, Stover AM, Riley WT, Cella D. Items banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and anger. in press. Manuscript in press.

Appendix

Listed below are the item stems for the recommended eight-item short forms for the PROMIS Pediatric Anxiety and Depressive Symptoms Scales. All items use a 7-day recall period (the preface is “In the past seven days”), and a 5-point response scale with the options *never* (0), *almost never* (1), *sometimes* (2), *often* (3) and *almost always* (4).

Anxiety:

I felt scared.

I worried about what could happen to me.

I felt worried.

I felt like something awful might happen.

I worried when I went to bed at night.

I thought about scary things.

I felt nervous.

I was afraid that I would make mistakes.

Depressive symptoms:

I felt like I couldn't do anything right.

I felt everything in my life went wrong.

I felt unhappy.

I felt lonely.

I felt sad.

I felt alone.

I thought that my life was bad.

I could not stop feeling sad.

Summed score to scale score translation for these short forms is in Table 7.

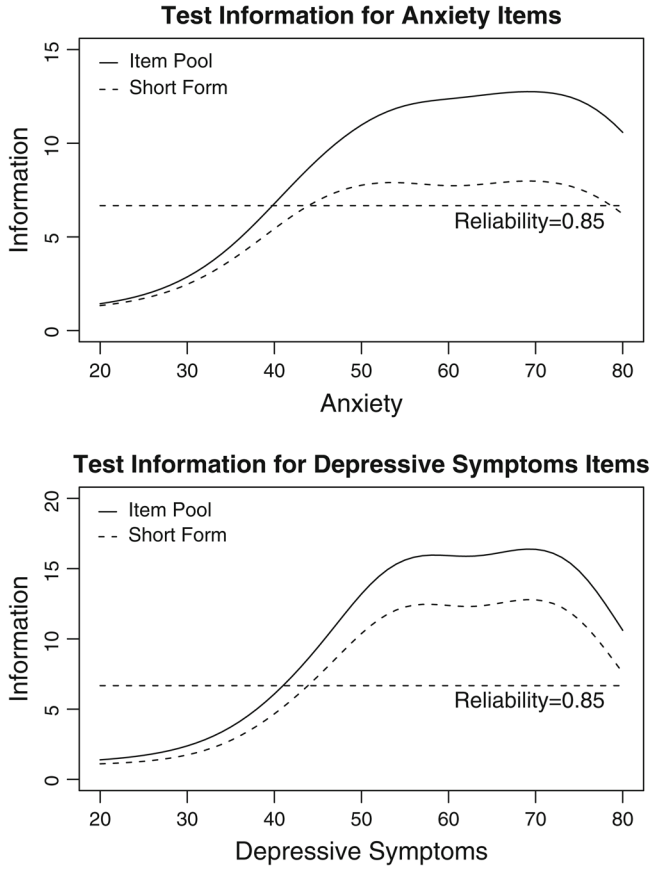

Fig. 1.

The *upper panel* shows the test information function for the anxiety item pool, and for the eight-item subset of the pool that is most informative near the middle of the distribution, and the *lower panel* shows the corresponding curves for the depressive symptoms item pool

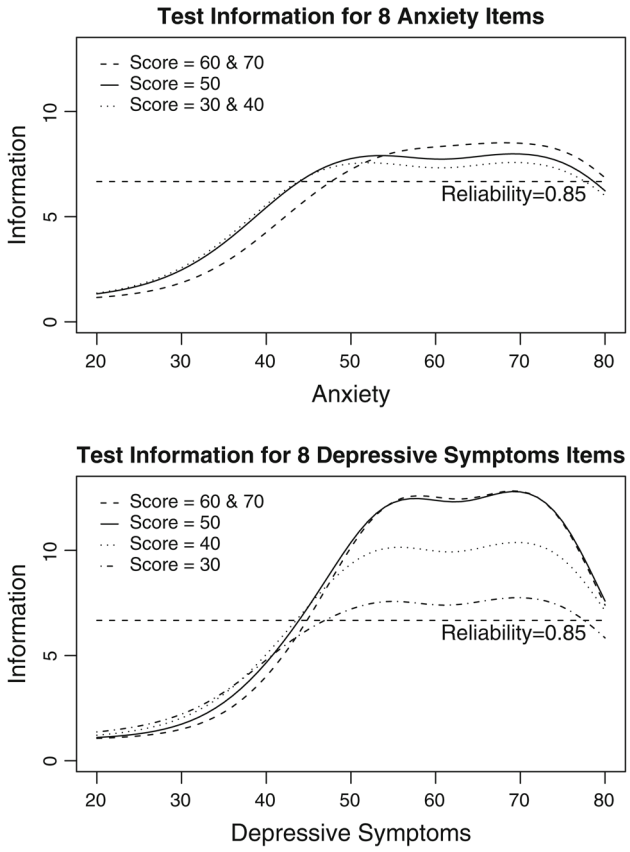

Fig. 2.

Test information curves for the most informative set of eight items from the anxiety (*upper panel*) and depressive symptoms (*lower panel*) item pools at *T*-scores of 30, 40, 50, 60, and 70

Table 1

PROMIS pediatric anxiety and depressive symptoms item stems administered on forms 1 and 2 for item analysis and calibration

	Form 1	Form 2
Anxiety	I got scared really easy.	I felt like something awful might happen.
	I felt afraid.	I felt scared.
	I worried about what could happen to me.	I was worried I might die.
	It was hard for me to stop worrying.	I worried when I went to bed at night.
	I woke up at night scared.	I worried when I was at home.
	I worried when I was away from home.	I thought about scary things.
	I was afraid that I would make mistakes.	I was afraid of going to school.
	I felt nervous.	I felt relaxed.
	It was hard for me to relax.	I felt worried.
Depressive symptoms	I wanted to be by myself.	I felt too sad to eat.
	I felt that no one loved me.	I didn’t care about anything.
	I cried more than usual.	I felt sad.
	I felt alone.	I felt lonely even when there were people with me.
	I felt like I couldn’t do anything right.	
	I felt so bad that I didn’t want to do anything.	I thought that my life was bad.
	I felt everything in my life went wrong.	It was hard for me to have fun.
	Being sad made it hard for me to do things with my friends.	I could not stop feeling sad.
		I felt stressed.
	It was hard to do school work because I felt sad.	I was bored.
		I felt lonely.
	I felt like crying.	I felt unhappy.

Table 2

Item calibration participants demographic and background information

	Form 1 *n* = 759 (%)	Form 2 *n* = 770 (%)	Total Form 1 and Form 2 *n* = 1,529 (%)
Child's gender			
Male	382 (50.3)	351 (45.6)	733 (47.9)
Female	377 (49.7)	419 (54.4)	796 (52.1)
Missing	0	0	0
Child's age (years)			
8–12	446 (58.8)	441 (56.4)	887 (57.7)
13–17	312 (41.1)	326 (42.3)	638 (42.0)
Missing	1 (0.1)	3 (0.3)	4 (0.3)
Child's race			
White	457 (60.2)	452 (58.7)	909 (59.4)
Black or African American	154 (20.2)	168 (21.8)	322 (21.1)
American Indian/Alaska native	5 (0.6)	10 (1.3)	15 (1.0)
Asian	12 (1.6)	13 (1.7)	25 (1.6)
Native Hawaiian/Pacific Is.	0	1 (0.1)	1 (0.1)
Other	58 (7.6)	50 (6.5)	108 (7.1)
Multiple races	47 (6.2)	54 (7.0)	101 (6.6)
Missing	26 (3.4)	22 (2.9)	48 (3.1)
Child's ethnicity			
Non-Hispanic	614 (80.9)	641 (83.2)	1255 (82.1)
Hispanic	141 (18.6)	121 (15.7)	262 (17.1)
Missing	4 (0.5)	8 (1.1)	12 (0.8)
Child's chronic conditions—6 months			
No	600 (79.0)	580 (75.3)	1180 (77.2)
Yes	157 (20.7)	187 (24.3)	344 (22.5)
Missing	2 (0.3)	3 (0.4)	5 (0.3)
Guardian's* relationship to child			
Parent	696 (91.7)	717 (93.1)	1413 (92.4)
Grandparent	32 (4.2)	30 (3.9)	62 (4.1)
Guardian or other	31 (4.1)	21 (2.7)	52 (3.4)
Missing	0	2 (0.3)	2 (0.1)
Guardian's* education level			
≤ 8th grade	12 (1.6)	16 (2.3)	28 (1.8)
Some high school	39 (5.1)	34 (4.4)	73 (4.8)
High school degree/GED	151 (19.9)	153 (19.7)	304 (19.9)
Some college/technical degree	255 (33.6)	245 (31.8)	500 (32.7)
College degree	179 (23.6)	214 (27.8)	393 (25.7)
Advanced degree	121 (15.9)	105 (13.6)	226 (14.8)
Missing	2 (0.3)	3 (0.4)	5 (0.3)

	Form 1 ***n*** **= 759 (%)**	**Form 2** ***n*** **= 770 (%)**	**Total Form 1 and Form 2** ***n*** **= 1,529 (%)**
Data collection site			
Schools—NC	57 (7.5)	57 (7.4)	144 (9.4)
Clinics—NC	349 (46.0)	350 (45.5)	699 (45.7)
Clinics—TX	353 (46.5)	363 (47.1)	716 (46.8)

* Guardian, parent, or care giver completing sociodemographic form and signing consent documents

Table 3

Factor loadings and residual correlations for an augmented bi-factor model fitted to the items on form 1

Item stem	General factor	Orthogonal group-specific factors			Doublet residual correlations
		Anxiety	Depressive symptoms	Afraid/scared	
I felt afraid.	0.68	*0.11*		0.67	
I got scared really easy.	0.64	0.30		0.38	
I felt afraid or scared.	0.76	0.15		0.20	
It was hard for me to stop worrying.	0.73	*0.11*			0.32
I worried about what could happen to me.	0.70	*0.10*			
I woke up at night scared.	0.72	0.40			
I was afraid that I would make mistakes.	0.70	−0.39			
It was hard for me to relax.	0.68	−0.15			
I worried when I was away from home.	0.64	*0.10*			
I felt nervous.	0.63	−0.19			
I felt everything in my life went wrong.	0.61		0.56		
I felt like I couldn’t do anything right.	0.62		0.55		
I felt so bad that I didn’t want to do anything.	0.57		0.48		
I felt alone.	0.59		0.50		0.27
I felt that no one loved me.	0.60		0.47		
Being sad made it hard for me to do things with my friends.	0.72		0.28		0.26
It was hard to do school work because I felt sad.	0.66		0.31		
I felt like crying.	0.66		0.24		0.49
I cried more than usual.	0.64		0.23		
I wanted to be by myself.	0.27		0.30		

Italicized entries are less than two standard errors from zero

Table 4

Factor loadings and residual correlations for an augmented bi-factor model fitted to the items on form 2

Item stem	General factor	Orthogonal group-specific factors		Doublet residual correlations
		Anxiety	Depressive symptoms	
I felt relaxed.	0.43	*0.03*		
I felt scared.	0.52	0.53		
I worried when I went to bed at night.	0.57	0.48		
I was worried I might die.	0.56	0.44		
I felt like something awful might happen.	0.59	0.41		
I worried when I was at home.	0.60	0.38		
I worry about what will happen to me.	0.68	0.33		
I thought about scary things.	0.56	0.32		
I was afraid of going to school.	0.58	0.19		
I felt worried.	0.65	0.37		0.36
I felt stressed.	0.58		0.36	
I felt lonely.	0.76		0.19	0.37
I felt lonely even when there were people with me.	0.74		*−0.03*	
I could not stop feeling sad.	0.84		−0.22	
I thought that my life was bad.	0.78		*0.04*	
I felt sad.	0.76		*−0.04*	
I felt unhappy.	0.76		0.23	
It was hard for me to have fun.	0.73		*0.07*	
I felt too sad to eat.	0.69		−0.27	
I didn’t care about anything.	0.47		*0.01*	
I was bored.	0.42		0.21	

Italicized entries are less than two standard errors from zero

Table 5

Item parameters and values for the *SS* X^2 fit index and LR DIF statistics for the anxiety items

Item stem	Item parameters					*SS* X^2 fit index			LR DIF		
	a	b_1	b_2	b_3	b_4	X^2	*df*	*p*	X^2	*df*	*p*
I felt scared.	1.89	−0.25	0.59	1.72	2.52	32	37	0.683	8.4	5	0.136
I worried about what could happen to me.	1.84	−0.24	0.48	1.54	2.21	49	39	0.134	6.5	5	0.261
I worried when I went to bed at night.	1.83	0.25	0.91	1.83	2.57	27	37	0.884	12.7	5	0.026
I felt worried.	1.81	−0.78	0.25	1.59	2.65	48	36	0.090	2.5	5	0.776
I felt like something awful might happen.	1.71	−0.43	0.51	1.75	2.65	44	37	0.207	10.7	5	0.058
I was worried I might die.	1.71	0.86	1.54	2.44	2.90	26	30	0.699	1.8	5	0.876
I woke up at night scared.	1.65	0.89	1.43	2.28	2.94	28	32	0.672	10.6	5	0.060
I worried when I was at home.	1.64	0.40	1.22	2.61	3.30	34	34	0.478	3.5	5	0.623
I felt nervous.	1.51	−0.85	0.18	1.86	2.85	58	37	0.016	8.8	5	0.117
I thought about scary things.	1.50	−0.40	0.51	1.85	2.64	28	37	0.846	11.6	5	0.041
I got scared really easy.	1.49	0.29	1.16	2.07	2.74	43	35	0.274	9.6	5	0.087
I was afraid that I would make mistakes.	1.48	−0.68	0.29	1.91	2.86	36	38	0.558	2.2	5	0.821
It was hard for me to relax.	1.42	−0.33	0.63	1.83	2.71	52	41	0.120	5.1	5	0.404
I worried when I was away from home.	1.32	0.77	1.50	2.59	3.16	29	36	0.785	2.6	5	0.761
I was afraid of going to school.	1.09	1.21	2.01	3.02	3.96	19	29	0.912	4.7	5	0.454
Items set aside due to LD											
I felt afraid.	1.61	0.26	1.18	2.31	3.31	33	35	0.559	7.6	5	0.180
It was hard for me to stop worrying.	1.87	0.06	0.85	1.76	2.52	24	35	0.930	3.5	5	0.623
Items set aside due to low discrimination											
I felt relaxed.	0.71	−1.07	1.26	3.59	4.80	39	39	0.468	5.1	5	0.404

The scale for the item parameters is set such that the distribution of anxiety in the reference population (represented by the NC portion of the sample) is standardized, mean 0 variance 1, as is conventional for reporting IRT parameters

Table 6

Item parameters and values for the *SS* X^2 fit index and LR DIF statistics for the depressive symptoms items

Item stem	Item parameters					*SS* X^2 fit index			LR DIF		
	a	b_1	b_2	b_3	b_4	X^2	*df*	*p*	X^2	*df*	*p*
I could not stop feeling sad.	2.53	0.61	1.13	1.92	2.46	25	32	0.804	5.8	5	0.326
I felt everything in my life went wrong.	2.46	0.35	0.96	1.74	2.19	29	34	0.694	10.9	5	0.053
I felt like I couldn't do anything right.	2.42	0.06	0.80	1.70	2.32	40	35	0.267	7.3	5	0.199
I felt unhappy.	2.14	−0.63	0.46	1.68	2.42	61	42	0.030	3.8	5	0.579
I felt alone.	2.11	0.31	0.98	1.91	2.58	44	34	0.122	10.4	5	0.065
I felt lonely.	2.04	−0.17	0.63	1.74	2.39	57	42	0.058	2.2	5	0.821
I thought that my life was bad.	2.00	0.25	0.77	1.80	2.41	58	40	0.033	5.3	5	0.380
I felt sad.	1.90	−0.75	0.27	1.74	2.75	59	41	0.035	3.3	5	0.654
Being sad made it hard for me to do things with my friends.	1.87	0.36	1.00	1.87	2.45	20	33	0.963	7.8	5	0.168
It was hard for me to have fun.	1.71	0.31	1.09	2.26	3.00	28	41	0.938	4.7	5	0.454
I felt too sad to eat.	1.45	1.02	1.70	2.62	3.41	25	33	0.842	12.9	5	0.024
I felt stressed.	1.27	−0.92	−0.02	1.54	2.61	49	48	0.446	5.9	5	0.316
I didn't care about anything.	1.03	0.05	1.12	2.65	3.65	50	46	0.316	4.7	5	0.454
I wanted to be by myself.	0.74	−1.88	−0.77	1.10	2.10	41	43	0.544	8.6	5	0.126
Items set aside due to LD											
I felt lonely even when there were people with me.	1.91	0.30	0.80	1.78	2.40	37	43	0.737	5.0	5	0.416
I felt that no one loved me.	2.08	0.69	1.20	2.08	2.54	19	31	0.950	11.2	5	0.048
It was hard to do school work because I felt sad.	1.90	0.52	1.03	1.89	2.43	32	35	0.600	13.7	5	0.018
Items set aside due to DIF											
I cried more than usual.	1.45	0.84	1.52	2.52	3.43	31	33	0.550	20.2	5	0.001
I felt so bad that I didn't want to do anything.	1.90	0.12	0.89	1.89	2.85	35	34	0.435	14.4	5	0.013
I felt like crying.	1.64	0.01	0.77	2.06	2.92	52	35	0.060	49.1	5	0.001
Items set aside due to low discrimination											
I was bored.	0.82	−3.05	−1.52	0.89	2.44	63	53	0.170	7.6	5	0.180

The scale for the item parameters is set such that the distribution of depressive symptoms in the reference population (represented by the NC portion of the sample) is standardized, mean 0 variance 1, as is conventional for reporting IRT parameters

Table 7

Summed score to scale score translation table for the recommended short forms

Summed score	Anxiety		Depressive symptoms	
	Scale score	SD	Scale score	SD
0	32	6	35	6
1	37	5	40	5
2	39	5	43	4
3	41	4	46	4
4	43	4	47	4
5	45	4	49	3
6	47	4	51	3
7	48	4	52	3
8	50	4	53	3
9	51	4	54	3
10	52	4	56	3
11	54	4	57	3
12	55	4	58	3
13	56	4	59	3
14	57	4	60	3
15	59	4	61	3
16	60	4	62	3
17	61	4	63	3
18	62	4	64	3
19	63	4	65	3
20	65	4	66	3
21	66	4	67	3
22	67	4	68	3
23	68	4	69	3
24	70	4	70	3
25	71	4	71	3
26	72	4	72	3
27	74	4	73	3
28	75	4	75	3
29	77	4	76	3
30	79	4	78	3
31	81	4	79	4
32	84	5	82	4

Scale scores are on a *T*-score scale; the values of SD are reported as conditional standard errors of measurement