Published in final edited form as:

Qual Life Res. 2017 March ; 26(3): 555–564. doi:10.1007/s11136-016-1467-3.

Impact of IRT item misfit on score estimates and severity classifications: an examination of PROMIS depression and pain interference item banks

Yue Zhao¹

¹Teaching and Learning Evaluation and Measurement Unit, The University of Hong Kong, Pokfulam, Hong Kong

Abstract

Purpose—In patient-reported outcome research that utilizes item response theory (IRT), using statistical significance tests to detect misfit is usually the focus of IRT model-data fit evaluations. However, such evaluations rarely address the impact/consequence of using misfitting items on the intended clinical applications. This study was designed to evaluate the impact of IRT item misfit on score estimates and severity classifications and to demonstrate a recommended process of model-fit evaluation.

Methods—Using secondary data sources collected from the Patient-Reported Outcome Measurement Information System (PROMIS) wave 1 testing phase, analyses were conducted based on PROMIS depression (28 items; 782 cases) and pain interference (41 items; 845 cases) item banks. The identification of misfitting items was assessed using Orlando and Thissen's summed-score item-fit statistics and graphical displays. The impact of misfit was evaluated according to the agreement of both IRT-derived *T*-scores and severity classifications between inclusion and exclusion of misfitting items.

Results—The examination of the presence and impact of misfit suggested that item misfit had a negligible impact on the *T*-score estimates and severity classifications with the general population sample in the PROMIS depression and pain interference item banks, implying that the impact of item misfit was insignificant.

Conclusions—Findings support the *T*-score estimates in the two item banks as robust against item misfit at both the group and individual levels and add confidence to the use of *T*-scores for severity diagnosis in the studied sample. Recommendations on approaches for identifying item misfit (statistical significance) and assessing the misfit impact (practical significance) are given.

Correspondence to: Yue Zhao.

Compliance with ethical standards

Conflict of interest The author has no conflicts of interest to declare.

Ethical approval The author declares that ethics approval is not required for this study because it uses anonymous secondary data sources.

Keywords

Item response theory; Model fit; Item misfit; Severity classifications; Patient-reported outcomes; PROMIS

Introduction

Item response theory (IRT) [1] has been increasingly applied in patient-reported outcome (PRO) measures in the last decade. An example is the use of IRT in developing and validating Patient-Reported Outcome Measurement Information System (PROMIS) item banks [2].

As an essential step in the IRT calibration process, the model-data fit is typically assessed by summarizing the discrepancy between observed values and the values expected under an IRT model using statistical tests of significance [3]. When significance tests suggest that an item does not fit well, the misfitting item is often removed from the item bank in PRO research. However, the misfitting item sometimes needs to be retained in the entire item bank due to its clinical relevance to the construct being measured or the unavailability of better fitting alternatives [4]. The question then becomes whether the observed misfit has an impact on the intended clinical applications, such as, estimating a patient's health status/trait and classifying severity levels for diagnosis (e.g., normal, mild, moderate or severe). Ultimately, it is the practical impact/consequence of misfit that should be considered when deciding on the merits of an IRT model for use in particular PRO situations. If the consequence is negligible, the misfit in the model may be tolerable [5]. This view is also echoed by the aphorism in statistics that “essentially all models are wrong but some are useful” [6].

Using statistical significance tests to detect misfit is usually the focus of IRT misfit evaluations, but such evaluations rarely address the consequences of using misfitting items and item statistics associated with them. A number of researchers [7, 8] have studied the practical consequences of misfit with various empirical educational testing datasets and concluded that the misfit was not always practically significant. However, in PRO research, which is distinct from educational testing settings, little is known about the extent to which the validity of inferences drawn from IRT-derived scores is robust against the misfit. As specified in the PROMIS psychometric evaluation and calibration plan [4], the issue is to determine to what degree misfit affects the model performance in terms of valid score interpretation. Moreover, for PRO measures that are used as screening and diagnostic tools, examining the impact of misfit on the score use for clinical diagnosis is of great clinical value and interest.

The main objective of this study was to evaluate the impact of IRT item misfit on score estimates and severity classifications. Analyses were conducted on the basis of item banks developed by the PROMIS [9] that measure depression and pain interference [10, 11]. This study also aims to demonstrate a recommended process of IRT model-fit evaluation, which not only limits to identifying misfitting items, but also involves assessing the impact of item misfit.

Methods

Sample and measure

The original PROMIS sample contained 21,133 participants, collected from the US general population and clinical populations in the PROMIS wave 1 testing phase [12], and their responses were collected from 14 domains including depression and pain interference. Two data collection designs were applied: full-bank design and block design. In the full-bank design ($N_{\text{full-bank}} = 7005$; all from the general population), respondents received all candidate items in a single domain. Full-bank cases responded to all items of a PROMIS domain along with all items of an existing legacy measure, which refers to a widely accepted and carefully validated measurement instrument for the respective domain. The full-bank design allowed for psychometric evaluation, such as, model-data fit and dimensionality, within a PROMIS domain. In the block design ($N_{\text{block}} = 14,128$; 6245 participants were from the general population, and 7883 were clinical participants), a block of 7 items in each of the 14 domains (98 items in total) was administered to respondents. The block design made it possible to examine the relationships across PROMIS domains. Details of sampling, item bank construction, and item content were documented elsewhere [4, 9–13]. Respondents were excluded from data analyses if they responded to less than half of the bank items or if their response time per item was less than a second on average and/or less than half a second for ten consecutive items.

Using secondary data sources collected from the aforementioned PROMIS wave 1 testing phase, analyses in the present study were conducted based on two PROMIS domains: depression and pain interference. In the domain of depression, 782 full-bank cases (all from the general population) completed the 28-item PROMIS depression item bank (PROMIS-DEP) along with the Center for Epidemiologic Studies Depression Scale (CES-D) [14], and 14,057 block cases (6213 general population cases and 7844 clinical cases) completed a subset of 7 items in the PROMIS-DEP. The 782 full-bank cases were used as the final sample for the current analyses, and the block cases were dropped for these analyses due to substantial missing responses. In the final depression sample ($n_{\text{PROMIS-DEP}} = 782$), 24% of this sample was scored 16 or higher and was indicated as at risk for clinical depression by the CES-D. The depression sample consisted of 375 males (48%) and 406 females (52%), ranging from 18 to 88 years old ($M = 51$ years, $SD = 19$). Seventy-eight percent of the respondents had completed higher than a high school level education.

In the pain interference domain, 845 full-bank cases (all from the general population) completed the 41-item PROMIS-pain interference item bank (PROMIS-PI) along with the Brief Pain Inventory Pain Interference (BPI-PI) subscale [15], and 14,003 block cases (6186 general population cases and 7817 clinical cases) completed a subset of seven items in the PROMIS-PI. The 845 full-bank cases were used as the final sample for the analyses in the current study, and the block cases were dropped for these analyses due to substantial missing responses. The pain interference final sample ($n_{\text{PROMIS-PI}} = 845$) included 387 males (46%) and 458 females (54%), ranging from 18 to 88 years old ($M = 51$ years, $SD = 18$). Eighty-three percent of the respondents had completed more than a high school level education.

Person scores that reflected a respondent's latent depression/pain interference severity levels were quantified by the PROMIS *T*-scores (mean = 50; standard deviation = 10), which are anchored to a general US population in terms of key demographics including age, gender, and race/ethnicity [12]. High *T*-scores reflected worse symptoms. Respondents' symptom severities on the *T*-score metric were classified into four categories: "normal," "mild," "moderate," and "severe." To identify the thresholds for differentiating each pair of adjacent severity categories, Cella et al. [16] conducted an empirical standard setting study in oncology and applied a modified bookmark method to derive these thresholds. In the study, clinical experts rank-ordered a series of vignette cards (showing the most likely response to each of five representative items in a PROMIS domain for a given *T*-score level) by severity, placed bookmarks to differentiate the cards into different severity categories and arrived at consensus. Cella et al. recommended severity thresholds differentiating each pair of adjacent severity categories from "normal" to "severe" as 55, 65, and 75 for the depression item bank and 50, 60, and 70 for the pain interference item bank; all were on the *T*-score metric. These thresholds were applied in the current study to classify the severity category for each respondent, with details described shortly.

Procedure

Prior to item calibration, IRT assumptions of unidimensionality and local independence were tested. To determine essential unidimensionality, a single-factor confirmatory factor analysis (CFA) model was employed based on the raw categorical data with a weighted least squares means and variance-adjusted estimator (WLSMV estimation) using Mplus 6 [17]. The goodness of fit of the CFA was evaluated using the comparative fit index (CFI), the Tucker Lewis index (TLI) and the root-mean-square error of approximation (RMSEA) [18–21]. A very good fit (CFI > .95; RMSEA < .05; TLI > .95) or an adequate fit (CFI > .90; RMSEA < .10; TLI > .90) from the single-factor CFA would support the unidimensionality assumption. Local independence was evaluated between item pairs using a residual correlation matrix produced by the single-factor CFA in Mplus 6. An absolute value in excess of .20 in the residual correlation [4] indicates possible local dependencies.

The evaluation of IRT model fit involved two steps: identifying misfitting items by using both statistical test and graphical approach, and then evaluating the impact of item misfit on *T*-score estimation and severity classifications.

The identification of misfitting items was assessed using Orlando and Thissen's [22] summed-score item-fit statistics ($S-X^2$) in IRTPRO [23], and misfit was indicated by significant results with a Benjamini and Hochberg [24] adjusted overall alpha level of .05.¹ In addition to statistical tests, the degree of IRT model-data fit was also assessed using graphical displays by plotting the residuals between the observed and IRT model-based expected frequencies at each summed score point for each score category. Misfit is indicated by the pattern of substantial data exhibiting a large departure from zero in the residual plot.

¹The overall alpha level of .05 was adjusted with the total number of items in the respective PROMIS item bank. The adjusted alpha values from the smallest to largest ranged from .0018 (.05/28) to .05 for the PROMIS-DEP and ranged from .0012 (.05/41) to .05 for the PRO-MIS-PI.

In each domain, item and person parameters based on all items in the full item bank were calibrated using the graded response model (GRM) [25]. After omitting the misfitting items from the respective item bank, the remaining items (the item set consisting of well-fitting items) were assessed in terms of the unidimensionality and local independence assumptions using the aforementioned procedure, and these items were calibrated using GRM. All calibrations were performed in MULTILOG 7.03 [26]. In each calibration, the Stocking and Lord characteristic curve (SL) method [27] was applied to place the newly calibrated PROMIS parameters into the established/published PROMIS metric [10, 11]. The STUIRT [28] computer program was used to apply the SL procedure. As a linear transformation method, the SL utilized the item parameter estimates of anchor items taken from the new calibration and the established PROMIS metric to determine the scaling constants (multiplicative and additive constants). In this case, the well-fitting item set was designated as the anchor items. Once the scaling constants were obtained, all item and person (θ) parameters from the new calibration were linearly transformed to the established PROMIS metric. In other words, the θ scores obtained from all calibrations were placed on the same metric, allowing for direct comparisons. The IRT-based θ scores were then converted to PROMIS T -scores ($T = [10 \times \theta] + 50$). In each domain, two T -scores were obtained for each respondent; T_1 was based on the calibration of all items, and T_2 was based on the calibration after omitting the misfitting items. Two severity classifications, SC_1 (based on the calibration of all items) and SC_2 (based on the calibration after omitting the misfitting items), were computed accordingly for each respondent by applying the corresponding severity thresholds (PROMIS-DEP: <55 normal, 55–64 mild, 65–74 moderate, and 75 severe; PROMIS-PI: <50 normal, 50–59 mild, 60–69 moderate, and 70 severe [16]).

The impact of misfit was evaluated in terms of the agreement of T -scores and severity classifications. The extent of these agreements serves as a measure for the practical significance of the observed item misfit. The T -score agreement was evaluated at both the group and individual levels. At the group level, Cohen's d , the root-mean-square difference (RMSD) and the Pearson product-moment correlation (r) of the T -scores between inclusion and exclusion of misfitting items were computed. Cohen's guidance on small, medium, and large effect sizes [29] was employed to judge whether the group mean difference was of practical significance. At the individual level, any absolute difference of the T -scores that was larger than the lower bound of minimally important differences (MID) ranges was considered to have practical significance. According to an empirical study conducted with PROMIS in advanced-stage cancer patients [30], the recommended T -score MID ranges were 3.0–4.5 for depression and 4.0–6.0 for pain interference. We thus considered 3 and 4 as the cutoffs, respectively, for the PROMIS-DEP and PROMIS-PI to examine whether the individual level T -score difference is of practical significance.

Regarding the agreement of severity classifications, the classification consistency, over-classification rate, and under-classification rate were calculated by averaging the percentages of respective respondents. A respondent was considered consistently classified if $SC_1 = SC_2$, over-classified if $SC_1 > SC_2$, and under-classified if $SC_1 < SC_2$.

Results

In both domains, the summed score distributions, Cronbach's alpha and adjusted item-total correlations were calculated based on all items in the full item bank and the item set excluding the misfitting items. The results are reported in Table 1. Summed score distributions were positively skewed for both domains, with skewness ranging from 1.30 to 1.84. The alpha coefficients were .98 for the PROMIS-DEP and .99 for the PROMIS-PI. The mean adjusted item-total correlation for the PROMIS-DEP was .79 and that for the PROMIS-PI was .81 (Table 1).

The IRT assumptions were tested. For the examination of dimensionality, the fit statistics demonstrated an adequate fit for both item banks (Table 1), which lent support to the essential unidimensional assumption. Local independence was largely assumed in both domains, with the exception of one item pair. In the PROMIS-PI item set after removing the misfitting items, the value of residual correlation between the item pair "How often did pain prevent you from sitting for more than 30 min?" and "How often did pain prevent you from sitting for more than 1 h?" ($r = .207$) was slightly greater than the threshold, which is likely due to the similar item content. In summary, the assumptions of unidimensionality and local independence were met to a great extent and hence we moved onto the IRT calibration.

In the misfit identification, both statistical tests and graphical displays were examined. Using $S-X^2$, one mis fitting item was detected in the PROMIS-DEP, and seven misfitting items were identified in the PROMIS-PI. The $S-X^2$ statistics are reported in Table 2. Taking one PROMIS-PI misfitting item ("How often did pain prevent you from walking more than 1 mile?") as an example, Fig. 1 illustrates the graphical presentation of assessing item fit by plotting the residuals/differences between the observed and expected frequencies at each summed score point for each score category. Residuals with great departures from zero (with an absolute value greater than 2) appeared in almost all score categories, providing evidence of misfit.

In each domain, after the IRT item calibration and SL scaling procedure, two T -scores were obtained for each respondent (on the same metric) to examine the agreement of T -scores between inclusion and exclusion of misfitting items. At the group level, Cohen's d effect size measures, the RMSD, and the r of the T -scores between inclusion and exclusion of misfitting items were computed. As seen in Table 3, the results suggested negligible mean differences and a nearly perfect agreement of the T -scores. At the individual level, any absolute difference in the T -scores that was larger than 3 for the PROMIS-DEP and 4 for the PROMIS-PI was considered practically significant. No cases were identified in the PROMIS-DEP, and one (.1%) case was reported in the PROMIS-PI, falling into the "normal" category. Taken together, the misfit impact on the T -score agreement was practically insignificant.

In addition to the magnitude of agreement, the direction is also noteworthy. When misfitting items were included, respondents' T -scores tended to be underestimated on average, as shown in the negative values of Cohens' d effect sizes. These findings are in accordance with the pattern of the severity parameter estimates (b 's). In both domains, the average b

parameter estimates tended to be higher when misfitting items were included. In other words, the misfitting items generally reflected a higher level of severity than the well-fitting items, which echoes the observation that the majority of respondents endorsed the lowest score category, while a few respondents endorsed the higher score categories for these misfitting items.

Furthermore, the practical significance of the observed item misfit was assessed in terms of the severity classifications. In both domains, classification consistency ranging from 98 to 99% implied very high agreement of severity classifications (Table 3). In the PROMIS-DEP, inclusion of misfitting items resulted in misclassification for six (.8%) respondents, four cases were under-classified, and two cases were over-classified (Table 3). In the PROMIS-PI, 18 respondents (2.1%) were misclassified when misfitting items were included. Nine cases were under-classified, and nine were over-classified (Table 3). Overall, the misfit impact on the severity classifications can be considered negligible because fewer respondents were affected; however, the misclassification rate for each severity category varied and ranged from 0 to 7.5% (Table 3).

Discussion and conclusions

The examination of the presence and impact of IRT item misfit suggested that item misfit had a negligible impact on the T -score estimates and severity classifications with the general population sample in the PROMIS depression and pain interference item banks, implying that the impact of item misfit was insignificant. The evidence reported in the study supports the T -score estimates in the two item banks as robust against item misfit at both the group and individual levels, and it adds confidence to the use of T -scores for severity diagnosis in the studied sample.

PROMIS item banks were employed as a demonstrated case in the present study to illustrate the recommended process of IRT model-fit evaluation, which involves identification of the presence of item misfit (statistical significance) and assessment of the impact of item misfit (practical significance). A detailed discussion follows, and recommendations are given.

Identification of the presence of item misfit (statistical significance)

In the identification of the presence of item misfit, we employed Orlando and Thissen's [22] summed-score item-fit statistics ($S-X^2$). As a summed score-based method, $S-X^2$ was found to perform better than the traditional Chi-square-type item-fit statistics as well as to yield acceptable power and maintain a nominal Type I error rate for detecting misfit with normally distributed data [31, 32]. However, the performance of $S-X^2$ in detecting misfit can be influenced by the data distribution and test length [32, 33]. While statistical tests to detect misfit can be sensitive to the aforementioned factors, another direction for assessing item fit is the use of graphical approaches, which is strongly advocated by Hambleton and Han [5]. Graphical approaches offer an effective and straightforward approach for researchers to examine the discrepancy between the observed and expected frequencies more closely at each score point. One simple graphical approach is to plot the observed and expected frequencies at each score point (either sum scores or latent trait estimates depending on the fit statistics) either in a residual plot (example in Fig. 1) or along item characteristic curves

(example in [8]). We highly recommend that both statistical tests and graphical presentations be considered when determining model-data misfit.

Assessment of the impact of item misfit (practical significance)

Researchers or practitioners who study model fit often discover a model-fit evaluation with statistical significance testing and ignore the step of assessing the practical significance of model misfit. Assessment of practical significance of misfit refers to an assessment of the extent to which inferences drawn based on the test scores may be robust against the model misfit. As equally, if not more, important as the analysis of statistical significance, assessing the practical significance of model misfit is a crucial step of model-fit evaluation; model fit or misfit can have consequences that one should consider when selecting an IRT model. In the present study, we demonstrated the distinction between practical significance and statistical significance through an example of two PROMIS item banks. Although the data suggested statistically significant item misfit, the nature of the observed misfit was of little practical consequence in terms of estimating scores and classifying severity categories. In other words, the impact of the lack of model-data fit may or may not be consequential for the major uses of the test scores.

The assessment of the practical significance of item misfit has emerged in the educational measurement literature [3, 7, 8], but there have been few, if any, attempts to assess the practical impact of IRT misfit on PRO datasets. Ideally, the misfit impact assessment would involve comparisons between decisions made from an operational model and the ideal decisions under a perfectly fitting model. Unfortunately, such assessment can rarely, if ever, be conducted in practice due to the lack of a perfectly fitting model. Sinharay and Haberman [7] suggested replacing the ideal decisions with decisions made from scores under a better fit scenario. Adapting their guidelines about the assessment of misfit impact for educational tests into the PRO context, our recommended approach involves the following process: The first step focuses on identifying misfit using statistical significance tests. If no misfit is detected as statistically significant, the scores derived from the operational IRT model can be used. If misfit is detected, one should move on to the next step to assess the practical significance of misfit. If the misfit observed is practically insignificant, the scores derived from the operational IRT model can be used. If the misfit is practically significant, one should consider improving the fit by, such as, using a better-fitting model or excluding the misfitting items and then using the scores derived from the improved fit scenario. This approach carries several advantages. First, it is easy to implement in most IRT software in practice and can easily be generalized to broader quality of life research contexts beyond PRO research. In our case, we considered exclusion of misfitting items as a means of improving fit. In other cases (e.g., in the test construction stage), the use of an alternative IRT model that yields superior fit may offer a better solution for improving the fit. Caution should be taken about the reliability of the remaining items after omitting misfitting items. When a considerable number of misfitting items is deleted, test score reliability is likely to be lowered due to the shortened test length.

In addition to presenting the procedure for assessing the impact of misfit, we utilized several indices as misfit impact measures. These measures involve two measurement levels: the

group level and the individual levels. As an effect-size-type measure focusing on the group level, Cohen's d calculated the differences of *group* means between inclusion and exclusion of misfitting items. At the individual level, every single respondent was examined through referring their T -score difference (resulting from misfit) to minimally important differences and comparing the agreement of individual severity classifications between with and without misfitting items. Additionally, it is worth highlighting that the minimally important differences and severity thresholds employed in the study were derived from empirical studies [16, 30] rather than as an arbitrary value, which enhances the applicability of the findings.

Limitations and future directions

Limitations of the study are acknowledged. We only utilized the full-bank sample in the IRT calibration and model-fit assessment with the intention of avoiding the influence of missing data due to the block design and reducing the sensitivity of Chi-square-type fit statistics to a large sample size. This, along with different computer programs, test lengths, and alpha levels used for detecting misfit may lead to discrepancies, wherein the misfitting items were identified relative to previous studies [10, 11]. Correspondingly, caution should be taken about the generalizability of the results to clinical samples. The findings of this study about the T -score performance are specific to the general population sample. With clinical respondents added in, the entire sample, including the general population and clinical respondents, would become more heterogeneous and less skewed. The performance of the item-fit statistics in detecting misfit and the impact of the observed misfit need to be further investigated; there is evidence from a recent simulation study [34] that different percentages of clinical respondents in a calibration sample containing both clinical and nonclinical cases resulted in noticeable differences in item parameter and latent trait estimates.

Additionally, this study examined the misfit impact in the context of the full item bank. Given the distinguishing feature of PROMIS short forms, either static [10, 11] or dynamic computerized adaptive testing (CAT) (e.g., [35]) forms, it would be of practical value to extend the misfit impact study from the full item bank to the short forms. Follow-up studies could be conducted to compare the PROMIS T -scores resulting from a short form with a given number of well-fitting items to T -scores resulting from another short form with the same length that contains some misfitting items. It is an on-going process to continuously collect evidence on the decisions made based on the PROMIS scores that can be used in various applications (e.g., short forms, linking across multiple PROs assessing a similar domain) to ensure valid score interpretation and usage when accounting for misfit. Lastly, the minimally important differences and severity thresholds used here were derived for cancer patients [14, 20] rather than general populations.

In conclusion, this study serves as an important start for assessing the practical impact/consequence of IRT item misfit in PRO research. We hope to draw attention for PRO researchers that, in the IRT model-fit assessment, not only identifying misfitting items (statistical significance), but also evaluating the misfit impact (practical significance) should be adopted as a common practice in IRT applications for PRO research and practice.

Acknowledgments

The author would like to thank PROMIS[®] Statistical Center for permission to use the data for this study. Any opinions, findings, and conclusions expressed in the paper are those of the author and do not necessarily reflect the views of the data provider. PROMIS[®] was funded with cooperative agreements from the National Institutes of Health (NIH) Common Fund Initiative (U54AR057951, U01AR052177, U54AR057943, U54AR057926, U01AR057948, U01AR052170, U01AR057954, U01AR052171, U01AR052181, U01AR057956, U01AR052158, U01AR057929, U01AR057936, U01AR052155, U01AR057971, U01AR057940, U01AR057967, U01AR052186). The contents of this article use data developed under PROMIS. These contents do not necessarily represent an endorsement by the US Federal Government or PROMIS. See www.nihpromis.org for additional information on the PROMIS initiative.

References

1. Hambleton, RK., Swaminathan, H., Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage; 1991.
2. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Hays RD, on behalf of the PROMIS Cooperative Group. Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*. 2010; 63:1179–1194. DOI: 10.1016/j.jclinepi.2010.04.011 [PubMed: 20685078]
3. Swaminathan, H., Hambleton, RK., Rogers, HJ. Assessing fit in item response models. In: Rao, CR., Sinharay, S., editors. *Handbook of statistics: Psychometrics*. London: Elsevier; 2007.
4. 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(5):S22–S31. DOI: 10.1097/01.mlr.0000250483.85507.04 [PubMed: 17443115]
5. Hambleton, RK., Han, N. Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In: Lenderking, WR., Revicki, D., editors. *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications*. Washington: Degnon Associates; 2005. p. 57–78.
6. Box, GEP., Draper, NR. *Empirical model building and response surfaces*. New York, NY: Wiley; 1987.
7. Sinharay S, Haberman SJ. How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*. 2014; 33(1):23–35. DOI: 10.1111/emip.12024
8. Zhao Y. Approaches for addressing the fit of item response theory models to educational test data. *Dissertation Abstract International*. 2008; 69:12A. (UMI No. 3337019).
9. 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(5 Suppl 1):S3–S11. DOI: 10.1097/01.mlr.0000258615.42478.55
10. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS[®]): Depression, anxiety, and anger. *Assessment*. 2011; 18(3):263–283. DOI: 10.1177/1073191111411667 [PubMed: 21697139]
11. Amtmann DA, Cook KF, Jensen MP, Chen WH, Choi SW, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. *Pain*. 2010; 150(1):173–182. DOI: 10.1016/j.pain.2010.04.025 [PubMed: 20554116]
12. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of Clinical Epidemiology*. 2010; 63(11):1169–1178. DOI: 10.1016/j.jclinepi.2009.11.021 [PubMed: 20688473]
13. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*. 2007; 45(5 Suppl 1):S12–S21. DOI: 10.1097/01.mlr.0000254567.79743.e2 [PubMed: 17443114]

14. Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*. 1977; 1(3):385–401. DOI: 10.1177/014662167700100306
15. Cleeland CS, Gonin R, Hatfield AK, Edmonson JH, Blum RH, Stewart JA, et al. Pain and its treatment in outpatients with metastatic cancer. *New England Journal of Medicine*. 1994; 330(9): 592–596. DOI: 10.1056/NEJM199403033300902 [PubMed: 7508092]
16. Cella D, Choi S, Garcia S, Cook KF, Rosenbloom S, Lai JS, et al. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research*. 2014; 23(10):2651–2661. DOI: 10.1007/s11136-014-0732-6 [PubMed: 24938431]
17. Muthén, LK., Muthén, BO. Mplus [Computer software]. Los Angeles, CA: Muthén & Muthén; 2006.
18. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*. 1980; 88(3):588. doi: 10.1037/0033-2909.88.3.588
19. Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sage Focus Editions*. 1993; 154:136. doi: 10.1177/0049124192021002005
20. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 1999; 6(1):1–55. DOI: 10.1080/10705519909540118
21. Lance CE, Butts MM, Michels LC. The sources of four commonly reported cutoff criteria what did they really say? *Organizational Research Methods*. 2006; 9(2):202–220. DOI: 10.1177/1094428105284919
22. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*. 2000; 24(1):50–64. DOI: 10.1177/01466216000241003
23. Cai, L., Thissen, D., du Toit, S. IRTPRO [Computer software]. Lincolnwood, IL: Scientific Software International; 2015.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*. 1995; 57:289–300. DOI: 10.2307/2346101
25. Samejima, F. Estimation of latent ability using a response pattern of graded scores. Chicago: Psychometric Society; 1969.
26. Thissen, D., Chen, WH., Bock, RD. Multilog 7.03 [Computer software]. Lincolnwood, IL: Scientific Software International; 2003.
27. Stocking ML, Lord FM. Developing a common metric in item response theory. *Applied Psychological Measurement*. 1983; 7(2):201–210. DOI: 10.1177/014662168300700208
28. Kim, S., Kolen, MJ. STUIRT: A computer program for scale transformation under unidimensional item response theory models (Version 1.0). Iowa Testing Programs, University of Iowa; 2004.
29. Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
30. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six PROMISCancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology*. 2011; 64(5):507–516. DOI: 10.1016/j.jclinepi.2010.11.018 [PubMed: 21447427]
31. Orlando M, Thissen D. Further investigation of the performance of S-X²: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*. 2003; 27(4): 289–298. DOI: 10.1177/0146621603027004004
32. Kang T, Chen T. Performance of the generalized S-X² item fit index for the graded response model. *Asia Pacific Education Review*. 2011; 12(1):89–96. DOI: 10.1007/s12564-010-9082-4
33. Kang T, Chen T. Performance of the generalized S-X² item fit index for polytomous IRT models. *Journal of Educational Measurement*. 2008; 45(4):391–406. DOI: 10.1111/j.1745-3984.2008.00071.x
34. Smits N. On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: A simulation study. *Quality of Life Research*. 2016; 25(7):1635–1644. DOI: 10.1007/s11136-015-1199-9 [PubMed: 26646807]
35. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*:

An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation. 2010; 19(1):125–136. DOI: 10.1007/s11136-009-9560-5

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

Example of residual plot for assessing item fit. *Note* For the item “How often did pain prevent you from walking more than 1 mile?” Orlando and Thissen’s summed-score item-fit statistics ($S-X^2$) is 222.77 ($p < .0001$). y -axis = residual, which is the difference between the observed frequency and the IRT model expected frequency at each summed score point for each score category. Zero cell frequencies are excluded from the plot

Results from classical item analysis and unidimensionality analysis for patient-reported outcome measurement information system (PROMIS) depression and pain interference item banks

Table 1

Scale	No. of items	Skewness of summed score	α^a	Adjusted (corrected for overlap) item-total correlation		Confirmatory factor analysis (CFA)		
				Range	Mean	CFI ^b	TLI ^c	RMSEA ^d
PROMIS depression (all items)	28	1.34	.98	.70–.86	.79	.98	.98	.065
PROMIS depression (without misfitting items)	27	1.30	.98	.72–.85	.79	.98	.98	.066
PROMIS-pain interference (all items)	41	1.84	.99	.61–.89	.81	.98	.97	.081
PROMIS-pain interference (without misfitting items)	34	1.83	.99	.60–.89	.81	.98	.97	.088

^a α Cronbach's alpha

^b CFI comparative fit index (CFI > .90 indicates adequate fit; CFI > .95 indicates very good fit)

^c TLI/Tucker Lewis index (TLI > .90 indicates adequate fit; TLI > .95 indicates very good fit)

^d RMSEA root-mean-square error of approximation (RMSEA < .10 indicates adequate fit; RMSEA < .05 indicates very good fit)

Table 2

Item-fit statistics for misfitting items

Item ^a	$S-X^2$ ^b	df	Probability ^c
<i>PROMIS depression item bank</i>			
I felt I had no reason for living	111.02	68	.0008
<i>PROMIS-pain interference item bank</i>			
How much did pain interfere with your ability to participate in leisure activities?	127.37	83	.0013
How much did pain interfere with your close personal relationships?	126.95	77	.0003
How much did pain interfere with your relationships with other people?	132.01	73	.0001
How much did pain interfere with your enjoyment of social activities?	115.15	70	.0005
How often did you avoid social activities because it might make you hurt more?	133.00	88	.0014
How often did pain prevent you from walking more than 1 mile?	222.77	134	.0001
How often did pain prevent you from sitting for more than 10 min?	119.29	76	.0011

^aOnly the items that were identified as misfit are reported in the table

^b $S-X^2$ Orlando and Thissen's summed-score item-fit statistics provided by IRTPRO

^cA significant result with a Benjamini–Hochberg adjusted overall alpha level of .05 is an indicator of item misfit

Table 3

Comparisons of T -scores and severity classifications for patient-reported outcome measurement information system (PROMIS) depression and pain interference item banks in the presence of misfit

Item bank	n^a	r^b	d^c	RMSD ^d	Classification consistency ^e (%)	Under-classification rate ^e (%)	Over-classification rate ^e (%)
Depression	782	.9999	-.0026	.1459	99.23	.51	.26
Pain interference	845	.9985	-.0091	.4081	97.87	1.07	1.07

Item bank	AD larger than MID ^f				Classification (all items)				Classification (without misfitting items)			
	Normal	Mild	Moderate	Severe	Normal	Mild	Moderate	Severe	Normal	Mild	Moderate	Severe
Depression n	0	0	0	0	590	150	37	5	591	146	40	5
%	0	0	0	0	75.45	19.18	4.73	.64	75.58	18.67	5.12	.64
Pain interference n	1	0	0	0	423	331	82	9	421	335	80	9
%	.12	0	0	0	50.06	39.17	9.70	1.07	49.82	39.64	9.47	1.07

Item bank	Under-classification			Over-classification		
	Mild	Moderate	Severe	Normal	Mild	Moderate
Depression n	1	3	0	2	0	0
% ^g	.68	7.50	0	0.34	0	0
Pain interference n	7	2	0	5	4	0
% ^g	2.09	2.50	0	1.19	1.19	0

^a n sample size

^b r Pearson product-moment correlation of T -scores between those obtained based on all items in the respective item bank and those estimated after omitting the misfitting item(s) in the same item bank

^c d Cohen's d of T -scores between those obtained based on all items in the respective item bank and those estimated after omitting the misfitting item(s) in the same item bank. Small effect if d is around .20, medium effect if d is around .50, and large effect if d is around .80

^d RMSD root-mean-square difference (RMSD) of T -scores between those obtained based on all items in the respective item bank and those estimated after omitting the misfitting item(s) in the same item bank

^e Respondents' estimated T -scores were classified into four categories "normal," "mild," "moderate," and "severe" by applying corresponding thresholds: depression: <55 normal, 55–64 mild, 65–74 moderate, and >75 severe; pain interference: <50 normal, 50–59 mild, 60–69 moderate, and >70 severe. Two severity classifications (SC1 and SC2) were computed for each respondent: SC1 was based on the calibration of all items, and SC2 was based on the calibration after omitting the misfitting items. A respondent was considered consistently classified if SC1 = SC2, over-classified if SC1 > SC2, and

under-classified if $SC1 < SC2$. The classification consistency, over-classification rate, and under-classification rate were calculated by averaging the percentages of respective respondents in the corresponding item bank

f_j Any absolute difference (AD) of T -scores at the individual level that is larger than the lower bound of minimally important difference (MID) of 3 for the depression item bank and 4 for the pain interference item bank is considered to be of practical significance

g %; the misclassification rate within each severity category was calculated as the number of misclassified cases in the respective severity category divided by the sample size of the same severity category based on the item set excluding misfitting items