



ORIGINAL ARTICLE

Application of the Patient-Reported Outcomes Measurement Information System (PROMIS) item parameters for Anxiety and Depression in the Netherlands

Jan van Bebbber^{1,2}  | Gerard Flens³ | Johanna T.W. Wigman^{1,2,4} | Edwin de Beurs^{3,5}  | Sjoerd Sytema^{1,4} | Lex Wunderink² | Rob R. Meijer⁶

¹Interdisciplinary Ctr Psychopathol and Emot Regulat, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

²Department of Education and Research, GGZ Friesland, Leeuwarden, The Netherlands

³Foundation for Benchmarking Mental Health Care, Bilthoven, The Netherlands

⁴Rob Giel Research Center (RGOc), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

⁵Department of Clinical Psychology, Leiden University, Leiden, Netherlands

⁶Department of Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands

Correspondence

Jan van Bebbber, ICPE (UMCG), PO Box 30001 (internal: CC72), NL-9700RB Groningen, The Netherlands.

Email: j.van.bebber@umcg.nl

Funding information

Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 016.156.019; Veni, Grant/Award Number: 016.156.019; Mental Health Care Center Friesland

Abstract

Objectives: The Patient-Reported Outcomes Measurement Information System (PROMIS) Health Organization has compiled and calibrated item banks for various domains in the United States, and these item banks have been translated into Dutch language.

Methods: The item banks for Anxiety and Depression have been administered in two samples, one drawn from the Dutch general and one drawn from the Dutch clinical population. The aim of this study was to investigate the appropriateness of the official PROMIS item parameters for these item banks that have been estimated based on data collected in the United States for use in the Netherlands. For both domains, we determined the fit of U.S. item parameters, the effect on individual domain scores and levels, the effect on correlations with full item bank totals, and the effect on classification accuracies of adaptive test scores for diagnoses of anxiety and mood disorders.

Results: The results showed that especially in the clinical population sample, fit appeared to be problematic for many items. However, simulations revealed that both sets of item parameters (official PROMIS vs. unique Dutch) perform nearly equally well in practice.

Conclusion: We tentatively conclude that the official PROMIS item parameters can be used for scaling respondents in the Netherlands.

KEYWORDS

anxiety, computerized adaptive tests (CATs), depression, PROMIS, Real Data Simulation (RDS)

1 | INTRODUCTION

1.1 | The Patient-Reported Outcomes Measurement Information System

From a patient's perspective, Patient-Reported Outcomes (PROs) such as the ability to carry out daily chores or the ability to participate in various social interactions are much more relevant than physical indicators and concepts of health, such as variability in heart rate, body mass

indexes, or functional magnetic resonance images. However, PROs are frequently not standardized across patient populations and studies, thereby limiting the comparability of scores across studies, and in addition, many PRO measures have low measurement precision (Bjorner, Kosinski, & Ware Jr, 2003; Juniper et al., 1996; Rector & Cohn, 1992).

In order to overcome these limitations, the Patient-Reported Outcomes Measurement Information System (PROMIS) research group collected candidate items for various patient reported outcomes in the United States (Cella et al., 2007; DeWalt, Rothrock, Yount, Stone,

and PROMIS Cooperative Group, 2007). Furthermore, data that were representative of the 2000 U.S. census were collected in the United States (Cella et al., 2010). Based on these data, final item banks were compiled. Item banks are collections of items that are all operationalizations of the same domain of interest. To indicate a respondent's level on these domains, the PROMIS Health Organization uses T-scores. That is, item banks are scaled in such a way that the resulting person scores first are standardized according to the 2000 U.S. census and are then rescaled to have a mean of 50 and a standard deviation of 10. To indicate a respondent's level on these domains, all PROMIS item banks are scored on this T-score metric.

For these collections of items, parameter values have been derived by means of item response theory (Embretson & Reise, 2013). More specifically, the Graded Response Model (Samejima, 1970) has been used. The parameter values can be used to compute IRT scale scores, to compile brief versions of questionnaires with optimal measurement properties for specific testing purposes (e.g., have maximum measurement precision for certain trait levels), and to enable computerized adaptive testing (CAT). In CAT, items that are presented to respondents are tailored to responses given to previous items. With each consecutive item, an updated person score is derived, and the item that increases measurement precision maximally for this score is utilized next. This process usually continues until a predefined measurement precision is reached. In CATs, fewer items are needed to derive reliable scores compared with assessments with traditional (fixed-length) questionnaires. For a more elaborate introduction to the topic of CAT, see Meijer and Nering (1999).

The aim of the PROMIS Health Organization is that these item banks will be used worldwide so that results from studies conducted in different countries can be compared more easily: "The main goal of the PROMIS initiative is to develop and evaluate, for the clinical research community, a set of publicly available, efficient and flexible measurements of PROs, including health-related quality of life (HRQL)" (Cella et al., 2010, p. 2). In addition, Terwee et al. (2014, p. 1734) "... expected that PROMIS will be implemented worldwide and that PROMIS instruments will experience rapid adoption, once their cross-cultural validity is documented". Data gathered in various countries with internationally accepted instruments could be more easily combined and reanalyzed in meta-analyses.

Recently, 17 PROMIS item banks for adults have been translated into Dutch language (Terwee et al., 2014). The adult PROMIS item banks for Anxiety and Depression were recently administered by the Foundation for Benchmarking Mental Health Care¹ in two samples, one stratified sample drawn from the Dutch general population and one convenience sample drawn from the Dutch clinical population (Flens, Smits, Terwee, Dekker, Huijbrechts, & de Beurs, 2017; Flens, Smits, Terwee, Dekker, Huijbrechts, Spinhoven, & de Beurs, 2017). This offers the opportunity to investigate whether item parameters are similar in the Dutch and the U.S. item banks. For reasons of simplicity, in the remainder of this article, we will refer to the item parameters that were derived in the United States as the PROMIS

item parameters and refer to those that were derived from data collected in the Netherlands as Dutch item parameters.

1.2 | Aims of this study

First, we investigated whether the PROMIS item parameters could also be used to describe the data sampled from the Dutch general and the Dutch clinical population. Second, we investigated the effect of using Dutch item parameters instead of using the PROMIS item parameters in simulated adaptive tests. In particular, we performed Real Data Simulations (RDS) using both parameter sets (a) to investigate differences in T-scores computed; (b) to investigate differences in levels of anxiety and depression, respectively, as proposed by Cella et al. (2014); (c) to compare the correlations of simulated adaptive test scores with unweighted full item bank total scores; and (iv) to compare the predictive power of simulated CAT scores for diagnoses of mood and anxiety disorders, respectively. Finally, we used the PROMIS item parameters to compare the distributions of anxiety and depressive symptom experiences across populations.

2 | METHODS

2.1 | Participants

The U.S. PROMIS Wave one data file (Cella et al., 2010) was used by Pilkonis et al. (2011) for estimating item parameters for the item banks Anxiety and Depression. For efficiency reasons, data were collected using a block design, where respondents did not have to respond to all items. As a result, approximately one third of the $N_{\min} = 2,243$ and $N_{\max} = 2,928$ (number of respondents in the block design varied across items) respondents responded to all items. About 100 of these cases were flagged due to unrealistically short response times and removed from further analyses (Pilkonis et al., 2011). In addition, respondents who answered less than 50% of the items from a specific domain were removed from further analyses for that specific domain. These criteria resulted in sample sizes of $N = 788$ and $N = 782$ participants for the PROMIS Anxiety and Depression samples, respectively. We used the item parameters calibrated using the block design and refer to them as the PROMIS item parameters.

The Dutch general population sample (Flens, Smits, Terwee, Dekker, Huijbrechts, & de Beurs, 2017; Flens, Smits, Terwee, Dekker, Huijbrechts, Spinhoven, & de Beurs, 2017) was obtained using an online panel (Desan Research Solutions; www.desan.nl). Respondents participated voluntarily in the panel and received a small financial compensation for participation. A sample of $N = 1,486$ respondents was drawn and stratified on gender, age, education level, ethnicity, and region. The response rate was 71% resulting in $N = 1,055$ respondents. Fifty-three respondents were excluded from further analyses because they showed suspicious response patterns (e.g., all responses in one category in combination with short response times of less than 10 min for all items of both item banks). The final general population sample consisted of $N = 1,002$ respondents. The composition of this sample represented the marginal composition of the Dutch general population in 2013 (Statistics Netherlands; www.cbs.nl) in terms of gender, age (younger, middle-aged, and older), education (low, middle, and high), ethnicity (Dutch natives, western, and nonwestern

¹The Foundation for Benchmarking Mental Health Care is a Dutch trusted third party, which aims to provide a country-wide performance benchmark to evaluate and compare treatment outcomes of mental health care providers in the Netherlands.

immigrants), and region (north, east, south, and west), with maximal deviations of 2.5% for each category. Detailed information on the stratification process can be found in Flens, Smits, Terwee, Dekker, Huijbrechts, & de Beurs (2017), Flens, Smits, Terwee, Dekker, Huijbrechts, Spinhoven, & de Beurs (2017).

For the Dutch clinical population sample, $N = 3,296$ patients with common mental disorders who started their treatment in ambulatory mental health care were invited by the Dutch mental health care provider Parnassia Group. Item banks were only administered when informed consent had been obtained. The patients' diagnoses (4th ed.; DSM-IV; American Psychiatric Association, 2000) were assessed prior to the study in two ways. First, a psychiatric nurse administered the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998) in Dutch (van Vliet & de Beurs, 2007) by phone. Second, the diagnoses were verified in interviews, and in case of comorbid diagnoses, the primary diagnosis was established. The response rate was 31% resulting in $N = 1,032$ & 24 patients were excluded from further analyses because of missing values on some items. The final clinical sample consisted of 1,008 patients. In terms of DSM-IV diagnoses, 44% had a primary diagnosis of mood disorder, 33% an anxiety disorder, and 23% a disorder not specified any further (e.g., attention deficit disorder, somatoform disorder, and personality disorder). For gender and age no systematic differences between nonresponders and responders were found (Flens, Smits, Terwee, Dekker, Huijbrechts, & de Beurs, 2017; Flens, Smits, Terwee, Dekker, Huijbrechts, Spinhoven, & de Beurs, 2017).

Extensive information on the demographic background of respondents in the four samples that were used in this study can be found in Table A1 in the supporting information of this article. The composition of the U.S. general population samples and of the Dutch general population sample was similar in terms of gender, age, and with respect to the percentage of respondents that attained a college degree. Respondents from the Dutch general population sample were somewhat less likely to have received an advanced degree compared with the U.S. general population samples. Furthermore, respondents in the Dutch clinical sample were approximately 12 years younger than respondents in the PROMIS wave-1 samples, and the Dutch clinical sample contains approximately 10% more females than the PROMIS wave one samples. Due to differences in the way in which ethnicity and relationship status were recorded in the United States and in the Netherlands, a more in-depth comparison of the four samples was not possible.

2.2 | Instruments

The selection of items for the PROMIS item banks for Anxiety and Depression has been thoroughly discussed in Cella et al. (2010). All items together with the official PROMIS item parameters can be found online (www.assessmentcenter.net). The items comprising the PROMIS Anxiety item bank can be found in Table A2.1, and the items comprising the PROMIS Depression item bank can be found in Table A2.2. These tables also list the labels that are used for convenience in the remainder of this article.

2.2.1 | Statistical analyses: Fit of item parameters

For each domain, we first ran one analysis in which we determined the fit of the PROMIS U.S. item parameters to the data of the Dutch

general population and Dutch clinical population sample.² This was done in IRTPRO (Cai, Du Toit, & Thissen, 2011) by entering the U.S. item parameters as starting values and setting the number of iterations of the Bock Atkinson Expectation Maximization algorithm equal to one. We used summed-score-based item diagnostics (Orlando & Thissen, 2000) to assess item-level fit. These test statistics can be used to evaluate differences between observed and expected (model implied) item score frequencies for various score levels. Score levels are summed scores without the item targeted in the specific item fit test. Note that for each combination of item bank and target population, nearly 30 tests are performed. Furthermore, with more than 1,000 respondents in each group, the tests of item fit are very powerful. These considerations led us to choose α overall to equal 0.01, resulting in a comparison-wise α of 0.0004 by the conventional Bonferroni correction as criterion indicating misfit. However, in our view, fit is best considered as a continuum and not as a dichotomy.

In order to get an idea of the magnitude of the effect of using the PROMIS item parameters instead of Dutch item parameters, we computed differences in expected item scores for 13 T-scores (from 30 to 90 with steps of 5) along the depression continuum using both parameter sets. Expected item scores are those item scores that are most likely, given the parameter values of items in combination with the theta-values that correspond to designated T-scores. We did this for those 23 items of the depression item bank that were also used in the study conducted by Cella et al. (2014).

2.2.2 | Statistical analyses: Real Data Simulations (RDS)

To evaluate the practical consequences of using the PROMIS item parameters that might not be optimal for scaling Dutch respondents, we used RDS (Sands, Waters, & McBride, 1997). RDS can be used to determine important characteristics of CATs that are not yet implemented in practice. All RDS were performed using the response patterns from the Dutch clinical population sample because the fit of the official PROMIS item parameters was much more problematic in this sample than in the Dutch general population sample (see Results section).

For each item bank, we ran two RDS.³ In the first run, we used the official PROMIS item parameters, and in the second run, we used item parameters that were calibrated using the data from both Dutch samples in Multiple Group Item Response Theory analyses (Flens, Smits, Terwee, Dekker, Huijbrechts, & de Beurs, 2017; Flens, Smits, Terwee, Dekker, Huijbrechts, Spinhoven, & de Beurs, 2017).

²We did not perform Differential Item Functioning (DIF) analyses because the official PROMIS item parameters have been calibrated in a block design for reasons of efficiency. Up to our knowledge, a combination of a blocked design with DIF analyses is not feasible. Additionally, DIF tests would take into account the estimation errors of the PROMIS parameter estimates, while in CAT applications, the true values of parameter estimates are assumed to be known. Thus, our fit tests are more stringent than DIF tests.

³The following settings have been used: The first item provided maximum information with respect to the group mean of the U.S. general population ($\theta = 0$). Furthermore, we used Expected A Posteriori as interitem estimator, combined with Minimum Expected Posterior Variance to select follow-up items. Four items were always administered. We used a standard error accompanying the person estimate of less than .45 ($\approx r_{xx} = .80$) as stopping rule.

First, we transformed all latent trait estimates to the T-score metric that is used by convention for all PROMIS item banks and computed differences in T-scores based on PROMIS item parameters and based on Dutch item parameters.

Second, we recoded these T-scores into the four (normal, mild, moderate, and severe) levels⁴ of anxiety and depression proposed by Cella et al. (2014) and computed differences between levels based on PROMIS versus Dutch item parameters.

Third, we used the adaptive test scores to compare the correlations of simulated adaptive test scores with unweighted item bank totals (simple total scores computed as sums of the scores of all items in a designated item bank).

In addition, for patients in the clinical sample, information on their current primary DSM-IV (American Psychiatric Association, 2000) diagnoses were available. We used this information to create two dummy variables. The first contrasted patients with and without anxiety disorder as primary diagnosis. The second dummy variable contrasted patients with and without any kind of mood disorder. Fourth, we compared the classification accuracies of CAT scores based on the aforementioned parameter sets (PROMIS and Dutch item parameters) for the DSM-IV diagnoses of having any kind of anxiety disorder and of having any kind of mood disorder. We used the program Firestar (Choi, 2009) to compile syntax to be used in R (R Core Team, 2014) to perform these analyses.

2.3 | The latent distributions of anxiety and depression in the Dutch general and Dutch clinical population

For each domain, we used the PROMIS item parameters to compute expected a posteriori IRT scale scores for respondents in the Dutch general population sample and in the Dutch clinical population sample. This was done to compare the distributions of anxiety and depressive symptom experiences in both Dutch samples to the distributions of anxiety and depressive symptom experiences in the U.S. general population. In the US general population, domain scores were fixated to have a mean T-score of 50 and a SD of 10.

3 | RESULTS

3.1 | Fit item parameters for the PROMIS Anxiety item bank

The results of the sum score-based item diagnostics for the 29 anxiety items for the Dutch general and Dutch clinical population samples can be found in Table A 3.1. According to the criterion of 0.0004 for significance, application of the PROMIS item parameters to the data from the Dutch general population resulted in acceptable fit for only nine out of 29 items. For the Dutch clinical population sample (columns five through seven), application of the PROMIS item parameters resulted in acceptable fit for only one item.

3.2 | Fit item parameters for the PROMIS

Depression item bank

The results of the summed-score based item diagnostics for the 28 PROMIS Depression items are displayed in Table A 3.2. In general, results were similar to those of the PROMIS Anxiety item bank. Application of the PROMIS item parameters to the data from the Dutch general population resulted in acceptable fit for nine out of 29 PROMIS Depression items. With respect to the Dutch clinical population sample (columns five through seven), only the response data to two items showed acceptable fit using the PROMIS item parameters.

In order to illustrate the procedure of the aforementioned sum score-based item diagnostics, observed and expected score frequencies for various score levels (total scores without the item targeted) on item EDDEP04, *I felt worthless*, in the Dutch general population sample are displayed in Table A4. We collapsed score levels in such a way as to create expected score frequencies of at least 100 for one response category. As can be seen, for nearly all score levels, much less respondents chose the lowest response option than the PROMIS item parameters predicted. With the exception of very high score levels, the reverse holds for the second and third response option.

In Table 1, differences in expected item scores using both parameter sets are displayed for the depression items conditional on 13 T-scores along the depression continuum. As can be seen, for most items and score levels expressed as of T-scores, usage of either PROMIS or Dutch item parameters led to the same expected item scores. The item for which we found most differences was item EDDEP04, *I felt worthless*.

3.3 | How serious is misfit for practical decisions? Results Real Data Simulations

The results of the comparisons of T-scores based on PROMIS versus Dutch item parameters are summarized in Table 2. For both item banks, application of PROMIS or Dutch item parameters led to absolute differences in individual T-scores of more than five points in approximately 12% of all cases. Differences of more than 10 points were found in 0.3% of all cases for the PROMIS Anxiety item bank and in 0.8% of all cases for the PROMIS Depression item bank.

In Table 3, the cross tabulation of levels of anxiety as proposed by Cella et al. (2014) based on PROMIS item parameters and levels of anxiety based on Dutch item parameters is displayed. The same cross tabulation for the Depression item bank may be found in Table A6. Differences of more than one level were only encountered two times, both for the depression item bank. Furthermore, for both item banks, both parameterizations led to the same levels of anxiety and depression in three out of four cases (78% for anxiety and 75% for depression).

When comparing the correlations between simulated adaptive test scores (in which either PROMIS or Dutch item parameters were used) and unweighted full item bank total scores, we found that the choice of PROMIS or Dutch item parameters had a small effect on the magnitudes of the correlation coefficients. Differences were very

⁴T < 55: Normal, 55–64.99: Mild, 65–74.99: Moderate, and T > 75: Severe.

TABLE 1 Differences in expected item scores caused by using Dutch item parameters instead of official PROMIS item parameters for 13 T-scores along the depression continuum

Item	T-score												
	30	35	40	45	50	55	60	65	70	75	80	85	90
I felt worthless						-1			1	2	1	1	1
I felt that I had nothing to look forward to													
I felt helpless										1			
I withdrew from other people								-1					
I felt that nothing could cheer me up										-1			
I felt that I was not as good as other people						-1		-1		-1			
I felt sad				1						-1			
I felt that I wanted to give up on everything						-1	-1						
I felt that I was to blame for things								1					
I felt like a failure						1				-1			
I had trouble feeling close to people													
I felt disappointed in myself					1								
I felt that I was not needed						-1							
I felt lonely						1				-1			
I felt depressed					1	1			1				
I felt discouraged about the future					1	1							
I found that things in my life were overwhelming					1	1		1		1			
I felt unhappy				1	1	1							
I felt I had no reason for living											1		
I felt hopeless						1	1			1			
I felt pessimistic					1				-1				
I felt that my life was empty						-1		-1					
I felt emotionally exhausted					1	1							

Note. PROMIS, Patient-Reported Outcomes Measurement Information System.

The blank spaces represent correspondence in item scores.

TABLE 2 Differences in T-scores based on official PROMIS item parameters and Dutch item parameters for the anxiety and depression item banks (cumulative percentages)

	DIFF > ABS (1)	DIFF > ABS (2)	DIFF > ABS (3)	DIFF > ABS (5)	DIFF > ABS (10)
Anxiety	71.1%	52.2%	31.2%	12.0%	0.3%
Depression	70.3%	51.2%	32.0%	12.6%	0.8%

Note. PROMIS, Patient-Reported Outcomes Measurement Information System.

TABLE 3 Crosstab levels of anxiety based on official PROMIS item parameters and based on Dutch item parameters

		Level Dutch item parameters				Total
		Normal	Mild	Moderate	Severe	
Level PROMIS item parameters	Normal	133	30	0	0	163
	Mild	19	273	32	0	324
	Moderate	0	108	344	11	463
	Severe	0	0	28	30	58
	Total	152	411	404	41	1,008

Note. PROMIS, Patient-Reported Outcomes Measurement Information System.

small, although when we used the Dutch item parameters, correlations were somewhat larger for both item banks. For the PROMIS Anxiety item bank, we found a correlation of $r = 0.921$ when using the PROMIS item parameters, whereas using the Dutch item parameters resulted in a correlation coefficient of $r = 0.932$. For the PROMIS Depression item bank, we obtained a correlation of $r = 0.925$ when

using the PROMIS item parameters, whereas the Dutch item parameters lead to a correlation of $r = 0.930$.

Three logistic regression analyses were conducted to predict whether respondents in the Dutch clinical population sample would suffer from an anxiety disorder. In the first analysis, the unweighted total scores of all PROMIS Anxiety items were used as predictor. In

the second analysis, the simulated adaptive test scores based on PROMIS item parameters were used as predictor, and in the third analysis, simulated adaptive test scores based on the Dutch item parameters were used as predictor. In all three analyses, the tests of full models against the constant only models were statistically nonsignificant, indicating that the test scores did not reliably distinguish patients with and without an anxiety disorder diagnosis, regardless of which item parameters (PROMIS or Dutch) were used. The constant only model for the dependent variable anxiety disorder diagnoses yielded a classification accuracy of 67.1% overall by predicting “no mood disorder” for every respondent. We also computed correlations between both sets of simulated adaptive test scores (one set based on Dutch item parameters and one set based on PROMIS item parameters). For anxiety, the correlation equaled 0.935, and for depression, the correlation was equal to 0.916. Note that since both coefficients are close to 1, the relative positions of individuals are roughly the same, independent of item parameters sets used.

Three additional logistic regression analyses were conducted to predict whether respondents in the Dutch clinical population sample would suffer from a mood disorder. The results of these analyses are displayed in Table 4.

The test of the first full model against a constant only model was statistically significant, indicating that the unweighted item bank total score distinguishes between respondents with and without a mood disorder diagnosis ($X^2 = 47.8$, $p < 0.01$ with $df = 1$; Nagelkerke's $R^2 = 0.062$). The test of the second full model against a constant only model was statistically significant, indicating that the simulated adaptive test score based on the PROMIS Depression item parameters distinguishes between respondents with and without a mood disorder diagnosis ($X^2 = 62.5$, $p < 0.01$ with $df = 1$; Nagelkerke's $R^2 = 0.081$). A test of the third full model against a constant only model was statistically significant, indicating that the simulated adaptive test score based on the Dutch Depression item parameters distinguished between respondents with and without a mood disorder diagnosis ($X^2 = 58.4$, $p < 0.01$ with $df = 1$; Nagelkerke's $R^2 = 0.076$).

The constant only model for the dependent variable mood disorder diagnoses yielded a classification accuracy of 59.4% overall. Both the CAT that was based on the official PROMIS item parameters and

the unweighted item bank totals increased the classification accuracy of the constant only model by 1.9 to 61.3%. Interestingly, the adaptive test scores that were based on Dutch item parameters increased the classification accuracy of the baseline model by 3 to 62.4%. All three models lead to only small increments in classification accuracies over the classification accuracy of the constant only model, a fact also expressed by the low values of Nagelkerke's R^2 .

3.4 | The latent distributions of anxiety and depression in the U.S. general population, the Dutch general population, and the Dutch clinical population

Table 5 displays the expected a posteriori means of the estimated scores and standard deviations for all three population samples in our study. Recall that the metrics of both domains have been fixed (identified) by setting both means equal to 50 and the standard deviations equal to 10 for the U.S. general population sample during calibration. Both means in the Dutch general population sample are very close to 50, and both standard deviations are close to 10. So in terms of both central tendency (operationalized by the means) and in terms of spread (operationalized by the standard deviations) of anxiety and depressive symptom experiences, the U.S. and the Dutch general populations are very much alike.

Not surprisingly, respondents in the Dutch clinical sample report much higher levels of anxiety ($M_{ANX.Dutch.Clinical} = 64.3$) and depressive symptom experiences ($M_{DEP.Dutch.Clinical} = 62.9$) on average than respondents in the general populations samples. Furthermore, the scores of respondents in the Dutch clinical population sample are

TABLE 5 Expected a posteriori means and standard deviations of posterior distributions based on official PROMIS item parameters in the T-score metric

Domain	Sample	Mean	SD
Anxiety	U.S. _{general}	50.0 ^a	10.0 ^a
	Dutch _{general}	49.9	10.1
	Dutch _{clinical}	64.3	8.6
Depression	U.S. _{general}	50.0 ^a	10.0 ^a
	Dutch _{general}	49.6	10.0
	Dutch _{clinical}	62.9	8.4

^aFixed during calibration.

TABLE 4 Logistic regression results for predicting mood disorder diagnosis

Variables	B	SE (B)	Wald X^2	Df	p	e^B	95% CI e^B
S_{DEP}^a	0.019	0.003	44.5	1	<0.01	1.019	1.013, 1.025
Model X^2	47.8						
N	1,008						
$CAT_{DEP-U.S.}^b$	0.646	0.087	54.8	1	<0.01	1.908	1.603, 2.270
Model X^2	62.5						
N	1,008						
$CAT_{DEP-Dutch}^c$	0.639	0.088	52.6	1	<0.01	1.895	1.589, 2.259
Model X^2	58.4						
n	1,008						

Note. PROMIS, Patient-Reported Outcomes Measurement Information System.

^aUnweighted item bank totals.

^bSimulated adaptive test scores using official U.S. PROMIS item parameters.

^cSimulated adaptive test scores using the Dutch item parameters.

more homogenous than the scores in both general population samples, as indicated by clearly lower standard deviations ($SD_{ANX,Dutch.Clinical} = 8.6$, $SD_{DEP,Dutch.Clinical} = 8.4$).

4 | DISCUSSION

4.1 | Summary of main findings

With respect to the Dutch clinical population, considering the results of the summed-score based item diagnostics, we found that the response data of very few items could be described sufficiently well by the PROMIS item parameters. With respect to the Dutch general population, only the response data for approximately one third of all PROMIS Anxiety and Depression items could be described reasonably well by the PROMIS item parameters. Interesting, however, was that using the PROMIS item parameters for all items of both item banks in RDS instead of the Dutch item parameters did not lead to substantial decrements in various indicators of validity.

At first glance, these two results may seem contradictory. But statistical significance (of misfit) does not imply practical significance, the latter referring to whether practical decisions (such as classifications of subjects) change due to misfit. As Sinharay and Haberman (2014) and Crişan, Tendeiro, and Meijer (2017) have shown, in many cases violations of model assumptions do not have much influence on practical decisions.

In addition, using the official PROMIS item parameters to compare the distributions of anxiety and depressive symptoms experiences across populations revealed that the samples of the general populations in the United States and in the Netherlands were quite comparable in terms of anxiety and depressive symptom experiences.

4.2 | Practical implications and recommendations

Although the fit statistics indicated that the PROMIS item parameters did not describe the Dutch data very well, especially for the Dutch clinical population sample, using the PROMIS item parameters instead of the Dutch item parameters did not lead to dramatic decreases in correlations and classification accuracies. Thus, for sake of simplicity and international comparability, for research purposes on group level, we recommend using the official PROMIS item parameters that have been calibrated in the United States by Pilkonis et al. (2011). For assessing individuals, however, the situation is more complex, and additional research is recommended (see below). Although most respondents received similar T-scores and the same severity levels, for both item banks, approximately 12% of all respondents showed differences in T-score larger than 5 and one fourth of all respondents were classified at somewhat different severity levels. Note that we cannot treat either scores (based on PROMIS or based on Dutch item parameters) as a gold standard, because both parameter sets performed moderately at best with respect to predicting which individuals did receive a diagnosis of anxiety or mood disorder. In addition, the predictive power of the simulated adaptive test scores based on the PROMIS Depression item bank was also weak. In our view, these observations cast doubt on the validity of both item banks for detecting cases of anxiety and depression in clinical populations.

4.3 | Strengths and limitations

To our knowledge, this is the first study that did not focus solely on fit indices when assessing the cross-cultural validity of measurement model parameter estimates but also incorporated various validity indices that are relevant for test practice.

One limitation of the study was that the procedure we used to compute fit statistics did not take into account the standard errors of the PROMIS item parameter estimates. Because approximately 2,000 respondents have been used in the original block design for calibrating the items, we assume that the accompanying standard errors were actually quite small, and thus, we expect that our results will not differ much from those we would have obtained when these standard errors had been incorporated. Another limitation of this study is the fact that the data in the United States have been collected 2006/2007, and the data in the Netherlands have been collected in 2014/2015. In addition to this, in the United States, the census of the year 2000 served as reference, and in the Netherlands, the composition of the Dutch general population in 2013 was used. The meaning of symptoms may change over the years, and these subtle changes may also affect item parameters.

Although the results with respect to prediction of diagnostic status are disappointing, we think that two remarks are important. First, all respondents in the clinical sample had received a DSM-IV diagnosis and all respondents were still in treatment for those disorders. In a sample without this restriction of range (e.g., including healthy controls), the predictor scores probably would have been more useful to discriminate respondents with an anxiety diagnosis from those without such a diagnosis. Furthermore, the PROMIS item banks were primarily developed for use in general populations.

4.4 | Directions for future research

To further investigate the validity of the PROMIS Anxiety and Depression item parameters for use in the Netherlands, we suggest the following: First, administer both item banks to respondents drawn from the Dutch general and Dutch clinical population, use RDS to compute simulated adaptive test scores according to both parameterizations, and determine for which test takers the severity levels differ. Second, ask these respondents and possibly also informed others (best friends and/or first-degree relatives), which severity levels best reflect the clients' conditions.

Furthermore, future research may investigate the fit of the official PROMIS item parameters for other PROMIS domains across different countries. This is also what the PROMIS Health Organization tries to accomplish by international research collaborations. But instead of performing numerous "pairwise" DIF analyses (United States versus a single foreign country), we advocate an approach that incorporates data collected in various countries in a single calibration study. If international comparability of scores is the core aim of the PROMIS Health Organization, efforts should be made to find parameter estimates that fit optimally in various countries where these parameters shall be implemented.

Another interesting direction for future research would be temporal invariance of the official PROMIS item parameter estimates,

because much research is longitudinal and not (only) cross-sectional. Are the item parameters invariant with respect to therapeutic interventions?

However, until item parameters may be based on truly international calibration samples, the existing official PROMIS item parameters may be implemented, even though results of strict fit tests seem to warn against their use.

ACKNOWLEDGEMENTS

This study was funded by a grant from the Mental Health Care Center Friesland, The Netherlands. J.T.W. Wigman was supported by Veni grant 016.156.019.

DECLARATION OF INTEREST STATEMENT

The authors have no competing interests.

ORCID

Jan van Bebber  <http://orcid.org/0000-0001-9453-1862>

Edwin de Beurs  <http://orcid.org/0000-0003-3832-8477>

REFERENCES

- American Psychiatric Association (2000). DSM-IV-TR: Diagnostic and statistical manual of mental disorders, text revision. Washington, DC: American Psychiatric Association, 75.
- Bjorner, J. B., Kosinski, M., & Ware, J. E. Jr. (2003). Using item response theory to calibrate the headache impact test (HIT™) to the metric of traditional headache scales. *Quality of Life Research*, 12(8), 981–1002.
- Cai, L., Du Toit, S., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software]*. Chicago, IL: Scientific Software International.
- Cella, D., Choi, S., Garcia, S., Cook, K. F., Rosenbloom, S., Lai, J., ... Gershon, R. (2014). Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research*, 23(10), 2651–2661.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... PROMIS Cooperative Group (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194. <https://doi.org/10.1016/j.jclinepi.2010.04.011>
- Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., ... PROMIS Cooperative Group (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, 45(5 Suppl 1), S3–S11. <https://doi.org/10.1097/01.mlr.0000258615.42478.55>
- Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*, 33(8), 644–645.
- Crîșan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, 41(6), 439–455. 0146621617695522
- DeWalt, D. A., Rothrock, N., Yount, S., Stone, A. A., & PROMIS Cooperative Group (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, 45(5 Suppl 1), S12–S21. <https://doi.org/10.1097/01.mlr.0000254567.79743.e2>
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.
- Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the Health Professions*, 40(1), 79–105. 0163278716684168.
- Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2017). Development of a Computerized Adaptive Test for Anxiety Based on the Dutch-Flemish Version of the PROMIS Item Bank. *Assessment*. Advance online publication. 1073191117746742.
- Juniper, E. F., Guyatt, G. H., Feeny, D. H., Ferrie, P., Griffith, L. E., & Townsend, M. (1996). Measuring quality of life in children with asthma. *Quality of Life Research*, 5(1), 35–46.
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, 23(3), 187–194.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64.
- Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS (R)): Depression, anxiety, and anger. *Assessment*, 18(3), 263–283. <https://doi.org/10.1177/1073191111411667>
- R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rector, T. S., & Cohn, J. N. (1992). Assessment of patient outcome with the minnesota living with heart failure questionnaire: Reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. pimobendan multicenter research group. *American Heart Journal*, 124(4), 1017–1025. doi:0002-8703(92)90986-6 [pii]
- Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 35(1), 139–139.
- Sands, W. A., Waters, B. K., & McBride, J. R. (1997). Computerized adaptive testing: From inquiry to operation. American Psychological Association.
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., ... Dunbar, G. C. (1998). The mini-international neuropsychiatric interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, 59(Suppl 20), 22–33. quiz 34–57
- Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, 33(1), 23–35.
- Terwee, C. B., Roorda, L. D., de Vet, H. C., Dekker, J., Westhovens, R., van Leeuwen, J., ... Boers, M. (2014). Dutch-flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Quality of Life Research: an International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 23(6), 1733–1741. <https://doi.org/10.1007/s11136-013-0611-6>
- van Vliet, I. M., & de Beurs, E. (2007). The MINI-international neuropsychiatric interview. A brief structured diagnostic psychiatric interview for DSM-IV en ICD-10 psychiatric disorders. [Het Mini Internationaal Neuropsychiatrisch Interview (MINI). Een kort gestructureerd diagnostisch psychiatrisch interview voor DSM-IV- en ICD-10-stoornissen]. *Tijdschrift voor Psychiatrie*, 49(6), 393–397. doi: TVPart_1639 [pii]

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: van Bebber J, Flens G, Wigman JTW, et al. Application of the Patient-Reported Outcomes Measurement Information System (PROMIS) item parameters for Anxiety and Depression in the Netherlands. *Int J Methods Psychiatr Res*. 2018;27:e1744. <https://doi.org/10.1002/mpr.1744>