Computerized adaptive testing in primary care: CATja

van Bebber, Jan

Chapter 5

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

This chapter was based on the manuscript:
van Bebber, J., Flens, G., Wigman, J.T.W., de Beurs, E., Sytema, S., Wunderink, L., and Meijer, R.R. (2018). *The Patient-Reported Outcomes Measurement Information System (PROMIS) item parameters for Anxiety and Depression: Applicability for the Dutch general and Dutch clinical population.*
Accepted for publication in International Journal for Methods in Psychiatric Research.

Abstract

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. Also, in earlier studies the item banks for Anxiety and Depression have been administered in two samples, one stratified sample drawn from the Dutch general population and one convenience sample drawn from the Dutch clinical population. The aim of this study was to investigate the validity of the official PROMIS item parameters for the item banks of Anxiety and Depression that have been estimated based on data collected in the United States for use in the Netherlands. For both domains, we determined (i) the fit of U.S. item parameters, (ii) the effect on individual domain scores and domain levels, (iii) whether using the official PROMIS item parameters instead of Dutch parameters would affect the magnitude of the correlations with full item bank totals, and, (iv) whether using the official PROMIS item parameters instead of Dutch parameters would affect the classification accuracies of adaptive test scores for diagnoses of anxiety- and mood disorders. The results showed that especially in the clinical population sample, fit appeared to be problematic for many items. However, simulations revealed that both types of item parameters perform nearly equally well in practice. We tentatively conclude that the official PROMIS item parameters can be used for scaling respondents in the Netherlands.

5.1 Introduction

5.1.1 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, the ability to participate in various social interactions, or the degree to which one experiences sleep disturbances are much more relevant than physical indicators and concepts of health, such as variability in heart rate, Body Mass Indexes, or (changes in) functional magnetic resonance images over time. However, PROs are frequently not standardized across patient populations and studies, thus limiting the comparability of scores across studies. Moreover, many PRO measures have low measurement precision (Cella et al., 2010).

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 U.S. (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 U.S. (Cella et al., 2010). Based on these data, final item banks were compiled. Item banks, or item pools, are collections of items that all pertain to the same domain or construct of interest. To indicate a respondent's level on these domains/constructs, 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 US census and are then rescaled to have a mean of 50 and a standard deviation of 10 by the well-known transformation $T = z * 10 + 50$.

For these collections of items, parameter values have been derived by means of item response theory (Embretson & Reise, 2013). These parameter values can be used (i) to compute IRT scale scores, (ii) 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 (iii) 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 to 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 the Dutch language (Terwee et al., 2014). Two of those, the adult PROMIS item banks for Anxiety and Depression, were recently administered by the Foundation for Benchmarking Mental Health Care[4] in two samples, one stratified sample drawn from the Dutch general population and one convenience sample drawn from the Dutch clinical population (Flens et al., 2017a, 2017b). This offers the opportunity to investigate whether the 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 U.S. as the PROMIS item parameters and refer to the item parameters that were derived from data collected in the Netherlands as Dutch item parameters. For research purposes, the official PROMIS item parameters are freely available upon request from the PROMIS Health Organization.

5.1.2 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 population and the Dutch clinical population. Second, we investigated the effect of using the PROMIS item parameters instead of Dutch item parameters in simulated adaptive tests. In particular, we performed Real Data Simulations (RDS) using both parameter sets (i) to investigate differences in T-scores computed, (ii) to investigate differences in levels of anxiety and depression respectively as proposed by Cella et al. (2014), (iii) 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.

[4] The 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.

5.2 Methods

5.2.1 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 emotional stability 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} = 2243 and N_{max} = 2928 (number of respondents in the block design varied across items) respondents in this block design responded to all emotional stability items. One hundred 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 (full item bank administrations, i.e. numbers of respondents that responded to all items from these item banks). For all analyses in this article, we used the item parameters calibrated in using the block design and refer to them as the PROMIS item parameters.

The Dutch general population sample (Flens et al., 2017a, 2017b) 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. Of these respondents, 53 respondents were excluded from further analyses because they showed suspicious response patterns (e.g., all responses in one category in combination with very short response times). 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 non-western immigrants), and region (north, east, south, and west), with deviations of maximal 2.5% for each category. Detailed information on the stratification process used can be found in Flens et al. (2017a, 2017b).

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 to respond to all items from the PROMIS Anxiety and Depression item banks online (Flens et al., 2017). In accordance with Parnassia's policy, item banks were only administered when informed consent had been obtained. The patients' diagnoses (4th ed.;

DSM–IV; American Psychiatric Association, 1994) 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 clinical face-to-face assessments and, in case of comorbid diagnoses, the primary diagnosis was established. The response rate in the clinical sample was 31% resulting in N = 1,032. Of these, 24 patients were excluded from further analyses because of missing values on some items. The final clinical sample thus 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, personality disorder). For the variables gender and age no systematic differences between non-responders and responders were found (Flens et al., 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 supplementary material 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 to the U.S. general population samples. Furthermore, respondents in the Dutch clinical sample were approximately twelve 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 demographic variables, such as ethnicity and relationship status, were recorded in the U.S. and in the Netherlands, a more in-depth comparison of the four samples was not possible.

5.2.2 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.assessmentcenternet). The items comprising the PROMIS Anxiety item bank can be found in Table A2.1 (appendix), and the items comprising the PROMIS Depression item bank can be found in Table A2.2 (appendix). These tables also list the labels that are used for convenience in the remainder of this article.

5.2.3 Statistical analyses: Fit of item parameters

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

clinical population sample[5]. This was done in IRTPRO (Cai, Du Toit, and 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 1000 respondents in each group, the tests of item fit are very powerful. These considerations led us to choose alfa overall to equal .01, resulting in a comparison-wise alfa of .0004 by the conventional Bonferroni correction as criterion indicating misfit. We note however, that 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 on the item level, we computed differences in expected item scores for thirteen 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).

5.2.4 Statistical analyses: Real Data Simulations

To evaluate the practical consequences of using the official PROMIS item parameters that might not be optimal for scaling Dutch respondents, we used *Real Data Simulations* (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[6]. 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

[5] Readers that are familiar with the framework of IRT might question why we did not perform Differential Item Functioning (DIF) analyses. We did not do so because the official PROMIS item parameters have been calibrated in a block design for reasons of efficiency, and up to our knowledge, a combination of a blocked design with DIF analyses is not feasible. In addition, DIF tests would take into account the estimation errors of the official PROMIS item parameter estimates, while in CAT applications, it is assumed that the true values of item parameter estimates would be known. That is, our fit tests are more stringent than DIF tests.

[6] The following settings have been used in the simulations: The first item was the one that provided maximum information with respect to the group mean of the U.S. general population ($\theta = 0$). Furthermore, we used Expected A Posteriori (EAP) as inter-item estimator, combined with Minimum Expected Posterior Variance

from both Dutch samples in Multiple Group Item Response Theory analyses (Flens et al., 2017a, 2017b).

First, we transformed all CAT scores to the for PROMIS item banks conventional T-score metric and computed the difference in T-scores based on PROMIS item parameters and based on Dutch item parameters for each item bank.

Second, we recoded these T-scores into the four (normal, mild, moderate, and severe) levels[7] 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 for each 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 (that is, generalized anxiety disorder, obsessive- compulsive disorder, specific phobia, social phobia, panic disorder with and without agoraphobia, or post-traumatic stress syndrome) as primary diagnosis. The second dummy variable contrasted patients with and without any kind of mood disorder (that is, first episode or recurrent depression, dysthymia, or depressive episode in bipolar disorder). Fourth, for each item bank, we compared the classification accuracies (count correct classifications divided by total count classifications) of CAT scores based on the aforementioned parameter sets (official PROMIS U.S. item parameters and item parameters estimated on Dutch data) 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.

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

For each domain, we used the official PROMIS item parameters to compute expected a posteriori (EAP) 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

(MEPV) to choose most appropriate follow-up items. A minimum of four items was always administered. When the standard error of the person estimate fell below .45, a value that corresponds to a reliability of .80, no more items were administered. We chose this cut-off value, because, according to the assessment criteria of the Dutch commission on test affairs (COTAN), a reliability of a least .80 is required to qualify an instrument as sufficiently reliable in contexts where important decisions about individuals' futures are made.

[7] T < 55: Normal, 55-64.99: Mild, 65-74.99: Moderate, and T > 75: Severe.

experiences in the U.S. general population. These scores were fixated to be standardized (M = 0 and SD = 1) in the PROMIS calibration samples, which served as points of reference.

5.3 Results

5.3.1 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 5.3.1 in the Appendix. According to the criterion of .0004 for significance, application of the official PROMIS item parameters to the data from the Dutch general population resulted in acceptable fit for only nine out of 29 anxiety items. For the Dutch clinical population sample (columns five through seven), application of the U.S. item parameters resulted in acceptable fit for only for one item according to our level of significance.

5.3.2 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 (Appendix). In general, results were similar to those of the PROMIS Anxiety item bank. Application of the official 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 (Table A 5.2.2, columns five through seven), only the response data to items EDDEP28 and EDDEP48 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 in the appendix. 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 from Table A4, 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 5.1, the differences in expected item scores using both parameter sets are displayed for the depression items conditional on thirteen T-scores along the depression continuum. As can be seen, for most items and score levels expressed in terms 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*.

Table 5.1 Differences in expected item scores caused by using Dutch item parameters instead of official PROMIS item parameters for thirteen T-scores along the depression continuum.

	T-score												
Item	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							

Blank spaces represent correspondence in item scores.

5.3.3 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 5.2. 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 ten 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.

Table 5.2 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 %

In Table 5.3, 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 A 5.6 (Appendix). Differences of more than one level were only encountered two times, both for the depression item bank. Furthermore, for both item banks, both parametrizations led to the same levels of anxiety and depression in three out of four cases (78% for anxiety and 75% for depression).

Table 5.3 Cross tabulation levels of anxiety based on official PROMIS item parameters and based on Dutch item parameters.

		Level Dutch item parameters				
		Normal	**Mild**	**Moderate**	**Severe**	**Total**
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	1008

When comparing the correlations between simulated adaptive test scores (in which PROMIS parameters or the Dutch 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 correlations coefficients. These differences were very small, although when we used the Dutch item parameters, the correlations were somewhat larger for both item banks. For the PROMIS Anxiety item bank, we found a correlation of r = .921 when using the PROMIS item parameters in RDS, whereas using the Dutch item parameters resulted in a correlation coefficient of r = .932. For the PROMIS Depression item bank, we obtained a correlation of r = .925 when using the PROMIS item parameters, whereas the Dutch item parameters lead to a correlation of r = .930. We also computed the 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 .935, and for depression, the correlation was equal to .916. Note that since both coefficients are close to one, the relative positions of individuals are roughly the same, independent of the item parameters used.

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 the PROMIS item parameters were used as predictor and in the third analysis, the 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 non-significant, 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 to simulate adaptive test scores. 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.

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 5.4.

Table 5.4 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}*	.019	.003	44.5	1	<.01	1.019	1.013,1.025
Model X^2	47.8						
N	1008						
$CAT_{DEP\text{-}U.S.}$**	.646	.087	54.8	1	<.01	1.908	1.603,2.270
Model X^2	62.5						
N	1008						
$CAT_{DEP\text{-}Dutch}$***	.639	.088	52.6	1	<.01	1.895	1.589,2.259
Model X^2	58.4						
N	1008						

*Unweighted item bank totals; **Simulated adaptive test scores using official U.S. PROMIS item parameters; ***Simulated adaptive test scores using the Dutch item parameters.

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<.01 with df = 1; Nagelkerke's R^2 = .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<.01 with df = 1; Nagelkerke's R^2 = .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 distinguishd between respondents with and without a mood disorder diagnosis (X^2 = 58.4, p<.01 with df = 1; Nagelkerke's R^2 = .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.

Note that although both types of adaptive test scores performed nearly equally well across all simulations, the Dutch item parameters were consistently slightly superior to the official PROMIS item parameters.

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

Table 5.5 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. Note that both means in the Dutch general population sample are very close to 50 and that 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.

Table 5.5 Expected a posteriori (EAP) means and standard deviations posterior distributions based on official PROMIS item parameters.

Domain	Sample	Mean	SD
Anxiety	$U.S._{general}$	50.0*	10.0*
	$Dutch_{general}$	49.9	10.1
	$Dutch_{clinical}$	64.3	8.6
Depression	$U.S._{general}$	50.0*	10.0*
	$Dutch_{general}$	49.6	10.0
	$Dutch_{clinical}$	62.9	8.4

*Fixed during calibration.

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 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).

5.4 Discussion

5.4.1 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 (one from the anxiety and two from the depression item bank) could be described sufficiently well by the official 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 official 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 Crisan, 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 U.S. and in the Netherlands were quite comparable in terms of anxiety and depressive symptom experiences.

5.4.2 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 U.S. by Pilkonis (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, and which did not. 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.

5.4.3 Strengths and limitations

To our knowledge, this is the first study that investigated the cross-cultural validity of the official PROMIS item parameters for the emotional stability item banks of Anxiety and Depression. Furthermore, it is one of the first studies 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 2000 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 U.S have been collected 2006/2007, while the data in the Netherlands have been collected in 2014/2015. In addition to this, in the U.S., the census of the year 2000 served as reference, while 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 from the general Dutch population), the predictor scores would have been more useful to better discriminate respondents with an anxiety diagnosis from those without such a diagnosis. Related to this is that the PROMIS item banks were primarily developed for use in the general population.

5.4.4 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 level best reflects the clients' situation.

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 (U.S 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? For example, does the construct of depression have the same meaning before and after recovery from a depressive episode?

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.

5.5 Appendix

Table A5.1 Demographic background of respondents in the four samples.

	$PROMIS_{ANX}$	**$PROMIS_{DEP}$**	**$DUTCH_{GEN}$**	**$DUTCH_{CLIN}$**
Sample size	788	782	1002	1008
Gender (% female)	52.0	51.9	52.1	61.6
Age – mean	51.0	51.0	48.9	38.4
Age – SD	18.9	18.8	16.5	13.0
College degree (in %)	18.0	18.1	18.8	---
Advanced degree* (in %)	13.1	12.9	9.3	---

* Master, Medical Doctor and PhD degree.

Table A5.2.1 Labels and items PROMIS item bank anxiety.

Label	Item
EDANX01	I felt fearful.
EDANX02	I felt frightened.
EDANX03	It scared me when I felt nervous.
EDANX05	I felt anxious.
EDANX07	I felt like I needed help for my anxiety.
EDANX08	I was concerned about my mental health.
EDANX12	I felt upset.
EDANX13	I had a racing or pounding heart.
EDANX16	I was anxious if my normal routine was disturbed.
EDANX18	I had sudden feelings of panic.
EDANX20	I was easily startled.
EDANX21	I had trouble paying attention.
EDANX24	I avoided public places or activities.
EDANX26	I felt fidgety.
EDANX27	I felt something awful would happen.
EDANX30	I felt worried.
EDANX33	I felt terrified.
EDANX37	I worried about other people's reactions to me.
EDANX40	I found it hard to focus on anything other than my anxiety.
EDANX41	My worries overwhelmed me.
EDANX44	I had twitching or trembling muscles.
EDANX46	I felt nervous.
EDANX47	I felt indecisive.
EDANX48	Many situations made me worry.
EDANX49	I had difficulty sleeping.
EDANX51	I had trouble relaxing.
EDANX53	I felt uneasy.
EDANX54	I felt tense.
EDANX55	I had difficulty calming down.

Table A5.2.2 Labels and items PROMIS item bank Depression.

Label	Item
EDDEP04	I felt worthless.
EDDEP05	I felt that I had nothing to look forward to.
EDDEP06	I felt helpless.
EDDEP07	I withdrew from other people.
EDDEP09	I felt that nothing could cheer me up.
EDDEP14	I felt that I was not as good as other people.
EDDEP17	I felt sad.
EDDEP19	I felt that I wanted to give up on everything.
EDDEP21	I felt that I was to blame for things.
EDDEP22	I felt like a failure.
EDDEP23	I had trouble feeling close to people.
EDDEP26	I felt disappointed in myself.
EDDEP27	I felt that I was not needed.
EDDEP28	I felt lonely.
EDDEP29	I felt depressed.
EDDEP30	I had trouble making decisions.
EDDEP31	I felt discouraged about the future.
EDDEP35	I found that things in my life were overwhelming.
EDDEP36	I felt unhappy.
EDDEP39	I felt I had no reason for living.
EDDEP41	I felt hopeless.
EDDEP42	I felt ignored by people.
EDDEP44	I felt upset for no reason.
EDDEP45	I felt that nothing was interesting.
EDDEP46	I felt pessimistic.
EDDEP48	I felt that my life was empty.
EDDEP50	I felt guilty.
EDDEP54	I felt emotionally exhausted.

Table A5.3.1 Summed score based item diagnostics for the PROMIS Anxiety items.

	$Dutch_{general}$			$Dutch_{clinical}$		
Label	**X^2**	**d.f.**	**P**	**X^2**	**d.f.**	**p**
EDANX01	643.78	104	.0001	689.92	163	.0001
EDANX02	252.19	98	.0001	288.53	162	.0001
EDANX03	512.06	106	.0001	348.26	175	.0001
EDANX05	245.29	122	.0001	531.99	155	.0001
EDANX07	202.24	109	.0001	768.10	170	.0001
EDANX08	189.30	133	.0010	806.27	193	.0001
EDANX12	376.49	124	.0001	497.46	160	.0001
EDANX13	262.84	158	.0001	618.33	223	.0001
EDANX16	260.63	158	.0001	463.70	227	.0001
EDANX18	120.65	111	.2498	296.24	175	.0001
EDANX20	168.59	158	.2673	410.48	234	.0001
EDANX21	218.02	145	.0001	406.96	198	.0001
EDANX24	226.16	168	.0019	329.89	241	.0001
EDANX26	320.54	152	.0001	835.12	214	.0001
EDANX27	173.97	131	.0071	779.32	192	.0001
EDANX30	670.55	129	.0001	263.27	158	.0001
EDANX33	328.89	84	.0001	255.67	163	.0001
EDANX37	203.54	167	.0283	384.23	238	.0001
EDANX40	182.84	94	.0001	766.76	145	.0001
EDANX41	261.92	107	.0001	542.57	166	.0001
EDANX44	241.71	175	.0006	263.03	246	.2174
EDANX46	199.42	118	.0001	294.51	148	.0001
EDANX47	308.90	135	.0001	731.12	186	.0001
EDANX48	279.60	133	.0001	623.77	164	.0001
EDANX49	312.36	198	.0001	599.52	235	.0001
EDANX51	161.60	158	.4054	266.75	189	.0002
EDANX53	200.00	114	.0001	609.50	146	.0001
EDANX54	194.89	121	.0001	263.68	146	.0001
EDANX55	171.24	120	.0015	311.76	168	.0001

Table A5.3.2 Summed score based item diagnostics for the PROMIS Depression items.

	$Dutch_{general}$			$Dutch_{clinical}$		
Label	**X^2**	**d.f.**	**P**	**X^2**	**d.f.**	**P**
EDDEP04	1822.90	99	.0001	648.09	148	.0001
EDDEP05	1044.77	110	.0001	327.08	163	.0001
EDDEP06	432.50	103	.0001	587.27	152	.0001
EDDEP07	321.73	144	.0001	348.24	194	.0001
EDDEP09	227.40	111	.0001	431.78	164	.0001
EDDEP14	245.73	151	.0001	390.27	223	.0001
EDDEP17	250.20	116	.0001	370.64	158	.0001
EDDEP19	379.23	115	.0001	332.97	182	.0001
EDDEP21	188.07	135	.0017	423.05	191	.0001
EDDEP22	198.45	106	.0001	586.14	159	.0001
EDDEP23	197.37	147	.0035	592.64	205	.0001
EDDEP26	231.35	131	.0001	274.13	174	.0001
EDDEP27	196.79	136	.0005	323.07	197	.0001
EDDEP28	170.30	154	.1746	248.45	204	.0183
EDDEP29	446.98	106	.0001	595.35	139	.0001
EDDEP30	331.88	130	.0001	712.14	192	.0001
EDDEP31	185.82	139	.0049	246.16	168	.0001
EDDEP35	210.60	131	.0001	850.74	185	.0001
EDDEP36	443.24	117	.0001	348.42	150	.0001
EDDEP39	331.40	94	.0001	285.01	172	.0001
EDDEP41	138.39	90	.0008	219.02	146	.0001
EDDEP42	168.00	147	.1132	446.98	205	.0001
EDDEP44	216.36	131	.0001	507.23	186	.0001
EDDEP45	151.17	129	.0886	338.77	187	.0001
EDDEP46	391.23	148	.0001	358.56	196	.0001
EDDEP48	129.43	131	.5228	245.75	184	.0016
EDDEP50	221.82	154	.0003	433.48	224	.0001
EDDEP54	289.64	150	.0001	409.48	192	.0001

Table A5.4 Observed and expected score frequencies for different score levels, Item EDDEP04, Dutch general population.

	Category 1		**Category 2**		**Category 3**		**Category 4**		**Category 5**	
Score level	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Obs.**	**Exp.**
0	134	134	0	0	0	0	0	0	0	0
1-2	93	106	13	0	0	0	0	0	0	0
3-6	81	113	25	2	9	0	0	0	0	0
7-13	75	114	38	8	10	1	0	0	0	0
14-22	47	102	51	27	32	4	2	0	0	0
23-44	28	78	105	102	78	40	11	3	0	0
45-96	3	5	21	31	83	83	54	44	9	8
97-112	0	0	0	0	0	0	0	0	0	0

Table A5.5 Content coverage of sets of fitting items.

Domain	Factor	Count item bank	Percentage item bank	Count subset	Percentage subset
Anxiety	1. Fear	7	.24	1	.13
	2. Anxious misery	11	.38	3	.38
	3. Hyperarousal	6	.21	3	.38
	4. Somatic symptoms	4	.14	1	.13
	5. Other	1	.03	0	.00
Depression	1. Negative mood	5	.18	0	.00
	2. Decreased positive affect	3	.11	2	.22
	3. Information processing deficits	3	.11	1	.11
	4. Negative views of the self	5	.18	1	.11
	5. Negative social cognition	4	.14	4	.44
	6. Other	8	.29	1	.11

Table A5.6 Crosstab levels of depression based on official PROMIS item parameters and based on Dutch item parameters.

		Level Dutch item parameters				
		Normal	**Mild**	**Moderate**	**Severe**	**Total**
Level PROMIS item parameters	**Normal**	138	41	1	1	181
	Mild	49	309	27	0	385
	Moderate	0	109	271	26	406
	Severe	0	0	6	30	36
	Total	187	459	305	57	1008

5.6 References

American Psychiatric Association, & American Psychiatric Association. (2000). DSM-IV-TR: Diagnostic and statistical manual of mental disorders, text revision. *Washington, DC: American Psychiatric Association, 75*.

Cai, L., Du Toit, S., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software]. *Chicago, IL: Scientific Software International.*

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.

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.

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33*(8), 644.

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), S12-21.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests*. Amsterdam: Nederlands Instituut van Psychologen.

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*.

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*.

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.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

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.*

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.