Development of a Computerized Adaptive Test for Anxiety Based on the Dutch–Flemish Version of the PROMIS Item Bank

Flens, G.; Smits, N.; Terwee, C.B.; Dekker, J.; Huijbrechts, I.; Spinhoven, P.; de Beurs, E.

DOI
10.1177/1073191117746742

Publication date
2019

Document Version
Final published version

Published in
Assessment

Article

Development of a Computerized Adaptive Test for Anxiety Based on the Dutch–Flemish Version of the PROMIS Item Bank

Assessment
2019, Vol. 26(7) 1362–1374

DOI: 10.1177/1073191117746742
journals.sagepub.com/home/asm
SAGE

Gerard Flens[1], Niels Smits[2], Caroline B. Terwee[3], Joost Dekker[3], Irma Huijbrechts[4], Philip Spinhoven[5], and Edwin de Beurs[1]

Abstract

We used the Dutch–Flemish version of the USA PROMIS adult V1.0 item bank for Anxiety as input for developing a computerized adaptive test (CAT) to measure the entire latent anxiety continuum. First, psychometric analysis of a combined clinical and general population sample (N = 2,010) showed that the 29-item bank has psychometric properties that are required for a CAT administration. Second, a post hoc CAT simulation showed efficient and highly precise measurement, with an average number of 8.64 items for the clinical sample, and 9.48 items for the general population sample. Furthermore, the accuracy of our CAT version was highly similar to that of the full item bank administration, both in final score estimates and in distinguishing clinical subjects from persons without a mental health disorder. We discuss the future directions and limitations of CAT development with the Dutch–Flemish version of the PROMIS Anxiety item bank.

Keywords

assessment, anxiety, clinical subjects, general population, item response theory, computerized adaptive test, PROMIS

In 2002, the National Institutes of Health started the patient-reported outcomes measurement information system (PROMIS) initiative in the Unites States of America (USA). PROMIS has the ambition to combine and transform all existing patient-reported outcome measures (PROMs) into one state-of-the-art assessment system for measuring self-reported health (Cella et al., 2007; Cella et al., 2010). With this system, self-reported health of adults and children is measured more accurately, precisely, responsively, and efficiently than existing PROMs allow for (Fries, Krishnan, Rose, Lingala, & Bruce, 2011; Fries, Rose, & Krishnan, 2011; Magasi et al., 2012; Pilkonis et al., 2014, Schalet et al., 2016). This is accomplished by the development of item banks (i.e., sets of items that measure the construct of interest) that meet high psychometric standards (i.e., good-quality item parameters). These item banks may be administered through a fixed questionnaire with a low number of items (also known as short forms), but preferably through a computerized adaptive test (CAT; Reeve et al., 2007). With short forms, the measurement precision for test outcomes can vary among respondents. A CAT, however, is more dynamic. It is a computer-administered test that selects questions based on the response pattern on previous questions until a precise outcome is obtained. In other words, it fixes the test outcomes' measurement precision and allows for the number of administered items to vary among respondents (Embretson & Reise, 2000). Consequently, administration burden can be reduced with a shorter test while maintaining the precision of the test result (Fliege et al., 2005).

PROMIS has become increasingly popular in the USA, and in other countries as well. By early 2017, many countries had developed translations of PROMIS item banks (http://www.nihpromis.org/measures/translations). Moreover, several countries had evaluated at least one item bank psychometrically (e.g., depression item bank: German, Jakob et al., 2015; Spanish, Vilagut et al., 2015). In the Netherlands, PROMIS is also gradually being implemented. First, 17 adult item banks and 9 pediatric item banks have been translated into Dutch–Flemish (Flemish is a variant of the Dutch language spoken in Belgium; Haverman et al., 2016; Terwee et al., 2014). Second, the item banks for Physical Function (Voshaar et al., 2014), Pain Interference (Crins et al., 2015), Pain Behavior (Crins et al., 2016), and

[1]Foundation for Benchmarking Mental Health Care, Bilthoven, Netherlands
[2]University of Amsterdam, Amsterdam, Netherlands
[3]VU University Medical Center, Amsterdam, Netherlands
[4]Parnassia Academy, Parnassia Psychiatric Institute, The Hague, Netherlands
[5]Leiden University, Leiden, Netherlands

Corresponding Author:
Gerard Flens, Stichting Benchmark GGZ (SBG), Rembrandtlaan 46, Bilthoven, Utrecht 3723 BK, the Netherlands.
Email: gerard.flens@sbggz.nl

Depression (Flens et al., 2017) have been psychometrically evaluated and meet the PROMIS standards (Reeve et al., 2007). Third, post hoc CAT simulations with the depression item bank have shown highly efficient and precise measurement for clinical subjects, with a similar accuracy compared with the full item bank administration (Flens et al., 2017).

Following depression, anxiety is the most common disorder in Dutch mental health care (de Graaf, ten Have, van Gool, & van Dorsselaer, 2012), and a worldwide problem in general (Baxter, Scott, Vos, & Whiteford, 2013). Validating the Anxiety item bank as input for a CAT administration is therefore an obvious next step before the PROMIS methodology can be implemented successfully in (Dutch) mental health care. New measurements that are more accurate, precise, responsive, and efficient are always desirable, but considering the nationwide implementation of routinely collected PROM data in the Netherlands, there is an urgent need for state-of-the-art efficient assessment with high-quality instruments (Carlier et al., 2012; de Beurs et al., 2011).

The present article has two goals. The first goal is to present a psychometric evaluation of the Dutch–Flemish version of the PROMIS adult V1.0 item bank for Anxiety (Pilkonis et al., 2011). The evaluation is conducted on a large sample with both clinical subjects and persons from the general population, because we aimed to develop an instrument that measures the full latent anxiety continuum (i.e., all persons with no symptoms of anxiety to patients with severe anxiety). Furthermore, the evaluation is based on the PROMIS standards to ensure high-quality items (Reeve et al., 2007), which is prerequisite for applying a CAT administration (Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012). Our second goal is to investigate how efficient and precise a CAT version of the Anxiety item bank may be to clinical and general population subjects, and how accurate this CAT version may be compared with a full item bank administration. For this goal, we performed a post hoc CAT simulation with a stopping rule set to a combination of high measurement precision and a fixed number of administered items. The stopping rule was chosen with a primary focus on the measurement precision of average and higher anxiety levels, as these are deemed the most relevant to measure, but without compromising the measurement precision of lower anxiety levels to a considerable extent. Efficiency and measurement precision were investigated both overall and as a function of the anxiety level; accuracy was investigated by comparing both test outcomes and group membership assignment between the CAT simulation and the full item bank administration.

Method

Participants

We collected data in a clinical and general population sample to cover the full range of possible latent anxiety levels in the Netherlands. For both samples, we aimed to include at least 1,000 respondents to obtain adequate item parameter estimates (Reise & Yu, 1990).

The eligible clinical sample consisted of 3,296 patients with common mental disorders who started their treatment in ambulatory mental health care. Patients were invited by the Dutch mental health care provider Parnassia Psychiatric Institute to digitally complete the item set. Parnassia Psychiatric Institute is by far the largest mental health institute in the Netherlands, and has a broad coverage across departments over the entire country. In accordance with the mental health care center's policy, the item set was only administered when written informed consent was obtained. The patient's diagnosis (*Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. [*DSM-IV*]; American Psychiatric Association, 1994) was assessed prior to the study in two steps. First, a psychiatric nurse administered the Dutch translation of the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998) by phone to ascertain the diagnosis. Second, the diagnosis was verified in a clinical face-to-face assessment.

The eligible general population sample consisted of 1,486 respondents that were approached digitally by a data collection panel to complete the item set (Desan Research Solutions; www.desan.nl). Respondents participated voluntary in the panel and received a small financial compensation for the study. To ensure representativeness of the sample, stratified sampling was applied. We used the following five stratification variables to mirror the Dutch population in 2013 (Statistics Netherlands; www.cbs.nl): gender (male, 49%; female, 51%), age (18-39 years, 34%; 40-64 years, 44%; 65+ years, 22%), education (low, 32%; middle, 40%; high, 28%), ethnicity (Dutch natives, 80%; Western immigrants, 10%; non-Western immigrants, 10%), and region (north, 10%; east, 21%; south, 22%; west, 47%). In each subgroup, deviations were allowed up to 2.5% because stratified sampling becomes increasingly difficult with an increasing number of variables. In addition, we assessed the diagnostic status of respondents by asking whether they were currently under treatment for mental health issues.

Measures

The item set consisted of 29 items from the Dutch–Flemish PROMIS adult V1.0 item bank for Anxiety (Terwee et al., 2014). The content of the items reflected a wide range of anxiety symptoms, problems, or negative affective states, and were stated positively (see Table 1; e.g., "I felt fearful"). Respondents were asked to indicate on a Likert scale how frequently they experienced the symptoms, problems or negative states in the past 7 days (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*), a higher score meaning more severe anxiety.

Table 1. IRT Item Characteristics for the Dutch–Flemish PROMIS Anxiety Item Bank Based on a Clinical Sample and General Population Sample.

Item code	Item	*M* (*SD*)	*a*	b_1	b_2	b_3	b_4	*H*	*S-X*^{2}	*p*
EDANX01	I felt fearful	2.45 (1.19)	2.75	0.03	0.78	1.74	3.13	0.70	340.09	.00***
EDANX02	I felt frightened	2.08 (1.10)	2.63	0.45	1.33	2.19	3.35	0.68	254.57	.14
EDANX03	It scared me when I felt nervous	2.25 (1.21)	2.51	0.38	1.05	1.92	3.09	0.67	343.54	.00***
EDANX05	I felt anxious	2.52 (1.25)	2.99	0.08	0.77	1.56	2.84	0.71	296.81	.00
EDANX07	I felt like I needed help for my anxiety	2.45 (1.39)	3.03	0.39	0.93	1.51	2.42	0.70	339.77	.00
EDANX08	I was concerned about my mental health	2.53 (1.38)	2.35	0.22	0.84	1.49	2.52	0.66	402.13	.00***
EDANX12	I felt upset	2.53 (1.22)	2.75	0.00	0.72	1.65	2.85	0.70	254.08	.19
EDANX13	I had a racing or pounding heart	2.31 (1.21)	1.80	0.19	0.99	2.00	3.37	0.60	396.79	.00
EDANX16	I was anxious if my normal routine was disturbed	2.21 (1.23)	2.20	0.43	1.17	1.95	3.08	0.64	362.27	.01
EDANX18	I had sudden feelings of panic	2.14 (1.24)	3.11	0.61	1.22	1.89	2.84	0.70	272.85	.05
EDANX20	I was easily startled	2.16 (1.12)	1.67	0.18	1.31	2.34	3.77	0.59	401.48	.00***
EDANX21	I had trouble paying attention	2.63 (1.19)	1.89	−0.38	0.63	1.65	3.15	0.63	327.94	.08
EDANX24	I avoided public places or activities	2.42 (1.30)	1.76	0.16	0.95	1.76	3.03	0.59	378.10	.12
EDANX26	I felt fidgety	2.80 (1.29)	2.99	−0.21	0.44	1.29	2.47	0.72	277.05	.06
EDANX27	I felt something awful would happen	2.04 (1.19)	2.28	0.65	1.37	2.14	3.06	0.64	404.70	.00***
EDANX30	I felt worried	3.08 (1.18)	2.35	−0.91	0.02	1.13	2.42	0.70	277.60	.04
EDANX33	I felt terrified	1.80 (1.06)	2.68	0.96	1.66	2.40	3.30	0.68	360.91	.00***
EDANX37	I worried about other people's reactions to me	2.51 (1.31)	1.97	0.02	0.85	1.65	2.70	0.62	400.32	.01
EDANX40	I found it hard to focus on anything other than my anxiety	2.32 (1.28)	3.59	0.40	1.04	1.71	2.64	0.72	244.75	.08
EDANX41	My worries overwhelmed me	2.47 (1.31)	2.87	0.22	0.87	1.61	2.59	0.70	326.91	.00
EDANX44	I had twitching or trembling muscles	1.96 (1.10)	1.36	0.66	1.58	2.75	4.53	0.52	370.35	.03
EDANX46	I felt nervous	2.59 (1.22)	2.87	−0.12	0.66	1.56	2.84	0.71	232.26	.36
EDANX47	I felt indecisive	2.36 (1.26)	2.26	0.25	0.95	1.80	2.97	0.65	378.12	.00
EDANX48	Many situations made me worry	2.52 (1.25)	2.85	0.08	0.74	1.58	2.84	0.70	278.73	.05
EDANX49	I had difficulty sleeping	2.73 (1.36)	1.34	−0.40	0.54	1.49	2.73	0.54	458.13	.00
EDANX51	I had trouble relaxing	2.90 (1.30)	2.51	−0.44	0.38	1.23	2.38	0.71	294.21	.12
EDANX53	I felt uneasy	2.52 (1.26)	3.06	0.07	0.80	1.57	2.72	0.72	282.94	.02
EDANX54	I felt tense	2.87 (1.32)	3.28	−0.27	0.42	1.21	2.28	0.74	267.90	.02
EDANX55	I had difficulty calming down	2.31 (1.24)	3.23	0.33	1.04	1.80	2.79	0.71	268.89	.02

Note. IRT = item response theory; PROMIS = patient-reported outcomes measurement information system; *N* = 2,010; *H* = Mokken's *H;* *S-X*^{2} = Orlando and Thissen's *S-X*^{2} statistic. Item code displays the original USA PROMIS item coding; *a* is the discrimination parameter; the *b*'s are threshold parameters; the item parameters are parametrized in the scale of the latent trait distribution of the general population (*M* = 0, *SD* = 1).
***$p < .001$.

Psychometric Evaluation

The psychometric evaluation of the Anxiety item bank was performed on the combined clinical and general population sample. We followed the PROMIS guidelines proposed by Reeve et al. (2007) to investigate whether we should remove any items from the item bank due to poor psychometric qualities. The evaluation focused on descriptive statistics, the main assumptions of item response theory (IRT), differential item functioning (DIF), and the item bank calibration. Below, we provide the details on these evaluation aspects. For more information, see Reeve et al. (2007). All statistical analyses were performed in the statistical environment R (R Core Team, 2015).

First, we evaluated the *descriptive statistics* of the full item bank sum scores (i.e., range, mean, standard deviation [*SD*], skewness, kurtosis, and internal consistency reliability [coefficient α]) and the individual item scores (i.e., response frequencies, range, mean, *SD*, skewness and kurtosis, interitem correlations, item-scale correlations, and drop in coefficient α for each item removed from the item bank). Specifically, undesirable patterns in the data were assessed (e.g., small range of item scores, outliers in item means, or negative correlations between items).

Second, we evaluated the IRT main assumptions of unidimensionality, local independence (LI), and monotonicity. *Unidimensionality* was evaluated with confirmatory factor analyses (CFA) using the R package lavaan (Version 0.5-18; Rosseel, 2012), and exploratory factor analyses (EFA)

using the R package psych (Version 1.5.4; Revelle, 2013), both conducted on the polychoric correlation matrix (Bollen, 1989). For CFA, we used the following (scaled) fit statistics to assess good fit of the one-dimensional model: comparative fit index (CFI) >0.95, Tucker–Lewis index (TLI) >0.95, root mean square error of approximation (RMSEA) <0.08, standardized root mean square residual (SRMR) <0.08 (Reeve et al., 2007). For EFA to indicate sufficient unidimensionality, the first extracted factor should explain above 20% of the variance (Reckase, 1979, as cited in Hambleton, 1988). Furthermore, the ratio of variance explained by the first to second factor should at least be 4 (Reeve et al., 2007).

The assumption of *LI* was evaluated with the residual correlation matrix from the single-factor CFA, and with Yen's Q3 statistic (Yen, 1993) using the R package mirt (Version 1.10; Chalmers, 2012). With the residual correlation matrix, we marked an item pair as possibly locally dependent when the corresponding coefficient was higher than 0.20 (Reeve et al., 2007). With Yen's Q3 statistic, the residual item scores are calculated under Samejima's graded response model (GRM; Samejima, 1969), and are then correlated among items. We assessed lack of model fit with Cohen's (1988) rules of thumb to interpret correlation effect sizes (Smits, Cuijpers, & van Straten, 2011): Q3 values between 0.24 and 0.36 imply moderate deviations of model fit, Q3 values above 0.37 imply large deviations. Item pairs with large deviations were marked as possibly locally dependent. When an item pair was marked by either it is residual correlation coefficient or Yen's Q3 statistic, further investigation was done by evaluating the impact of each item on the item parameter estimates (Reeve et al., 2007). To study this impact, we compared the item parameter estimates of the original GRM with a restricted GRM (i.e., minus one item).

The assumption of *monotonicity* was evaluated by examining graphs of item mean scores as a function of rest scores (total raw score minus the item score) using the R package Mokken (Version 2.7.7; van der Ark, 2007). In addition, we evaluated the accompanying scalability coefficients (Mokken's *H*) for the full scale and the individual items. Mokken's *H* was interpreted as follows: low quality when $.30 \leq H < .40$, moderate quality when $.40 \leq H < .50$, and high quality when $H \geq .50$ (Mokken, 1971).

Third, we evaluated uniform and nonuniform DIF (Embretson & Reise, 2000) for gender, age (recoded into a binary variable by means of a median split), and education level (low, medium, high). Both types of DIF were assessed with ordinal logistic regression (OLR) methods (Crane, Gibbons, Jolley, & van Belle, 2006) using the R package lordif (Version 0.2-2; Choi, Gibbons, & Crane, 2011). As measure of effect size, we used the change in McFadden's pseudo R^2, following the suggestion of .02 as critical value for rejecting the hypothesis of no DIF (Choi, Gibbons, et al., 2011).

Last, we estimated the item parameters of the Anxiety item bank (*calibration*) under the normal GRM (Samejima, 1969), an IRT model for polytomous items (Reeve et al., 2007). The GRM was fitted with multiple group estimation (McDonald, 1999; Smits, 2015) using the R package mirt (Version 1.10; Chalmers, 2012). We specified population (clinical and general) as grouping factor, and fixed the item parameters to be equal across groups. The latent trait (θ) was standardized to a scale with a mean of 0 and a standard deviation of 1 for the general population, a higher θ meaning more severe anxiety. The mean and standard deviation of the clinical sample were estimated under the model. As estimation algorithm, we used expectation–maximization. This algorithm is effective with one to three factors (Chalmers, 2012).

We evaluated the fit of the GRM by examining the item parameters and item fit. The GRM uses two types of parameters: the discrimination parameter *a* expresses the extent to which persons with similar θ estimates can be differentiated by the item; the four threshold parameters b_1 to b_4 (the number of threshold parameters for an item is equal to the number of response categories minus one) express the values of θ on which a person is expected to choose a higher over a lower item response. In addition, item fit was examined with the $S\text{-}X^2$ statistic (Orlando & Thissen, 2000, 2003). This statistic compares the observed and expected response frequencies under the used IRT model, and quantifies differences between these frequencies. Items with a $S\text{-}X^2$ $p < .001$ are considered to have a poor fit in the IRT model (Reeve et al., 2007). To study the impact of poor fit, we evaluated the effect of each item on the item parameter estimates by comparing those of the original GRM with those of a restricted GRM (i.e., minus one item).

Finally, we evaluated how well the item bank could measure Anxiety for the full latent continuum. To accomplish this, we plotted the test information of the item bank for $-4 \leq \theta \leq 4$. It is calculated as the sum of all item information values at any relevant θ level.

CAT Simulation

We used a post hoc CAT simulation to assess how efficient and precise a CAT version of the Anxiety item bank may be in clinical and general population subjects, and how accurate this CAT version may be compared with a full item bank administration. Previous studies have shown that post hoc CAT simulations are useful for this purpose as the results tend to be very similar to that of a real CAT administration (Kocalevent et al., 2009). Below, we provide the details on the CAT simulation settings and the assessment of efficiency, precision, and accuracy. The CAT simulation was performed using the R package mirtCAT (Version 0.5; Chalmers, 2015).

A CAT administration/simulation consists of four basic building blocks: a starting item, a method for estimating θ, an item selection procedure, and a stopping rule. The administration/simulation starts by presenting a first item. After a response is given, the software estimates θ and calculates the corresponding measurement precision (standard error [*SE*]). It then evaluates whether the obtained results meet the stopping rule. If not, a new item is selected and the procedure is repeated until the stopping rule is met, or all items have been presented.

As starting item, the CAT simulation used the item with the highest Fisher's information (Embretson & Reise, 2000; Wainer et al., 2000) at the average value of the latent trait in the general population (θ = 0). This item was *I felt tense*, which was coded as EDANX54 (Emotional Distress—ANXiety item bank, item 54) in the original USA PROMIS item bank (https://www.assessmentcenter.net).

To estimate θ, we could choose from two methods: maximum likelihood (ML) and Bayesian estimation (Embretson & Reise, 2000). Bayesian estimation is often chosen because it uses an a priori population distribution of the latent variable. This property ensures that θ can be estimated for all response patterns. A drawback of Bayesian estimation, however, is that the estimation of θ is also influenced by the a priori distribution; it pulls θ estimates toward the center of the population distribution, which may result in bias (Flens et al., 2017; Smits, 2015). ML, by contrast, does not use an a priori distribution, and is therefore not able to estimate θ for response patterns that exclusively comprise extreme responses. It is, however, a more stable estimator considering possible bias. ML can also result in bias, but generally to a lesser extent compared with Bayesian estimation, especially using CAT (Wang & Vispoel, 1998). Bias in ML emerges when the respondent's latent trait level is different from the average threshold of the administered items. This means that, under the assumption that the item bank has an adequate number of items to cover the entire latent continuum, bias under CAT should be minimal, as it is specifically designed to select items according to the threshold level at the provisional θ estimate. Consequently, we have chosen to use ML as the method for estimating θ. To deal with the issue of estimating θ for response patterns that exclusively comprise extreme responses, we could either set scale boundaries (Kim, Moses, & Yoo, 2015) or temporarily use a different estimation method (Chalmers, 2015). Due to a certain randomness in setting scale boundaries, we chose to temporarily use the commonly adopted Bayesian estimation method maximum a posteriori (Embretson & Reise, 2000). Thus, maximum a posteriori was used to estimate θ for response patterns that only include item scores 1 or 5, ML was used to estimate θ for all other response patterns.

To select additional items, we again used Fisher's information (Embretson & Reise, 2000; Wainer et al., 2000). Consequently, the item which had the highest information at the provisional θ estimate was selected.

As stopping rule, several methods have been proposed: a fixed number of administered items, a prespecified level of *SE*(θ), a prespecified change in θ estimate, or a prespecified change in *SE*(θ) (Babcock & Weiss, 2013; Choi, Grady, & Dodd, 2011; Smits et al., 2012). Each of these methods can be used individually or combined with each other. For this study, we chose to combine a prespecified level of *SE*(θ) with a fixed number of administered items. This combination rule is useful for measurements that are developed for both clinical and general population subjects. While clinical subjects mostly result in highly precise measurement with a low number of administered items (Flens, Smits, Carlier, van Hemert, & de Beurs, 2016), general population subjects often do not, not even when the full item bank is administered (Flens et al., 2017). Including a fixed number of administered items in the stopping rule should therefore result in efficient measurement for general population subjects as well, but without compromising the *SE* substantially.

The combination rule that we used to terminate the CAT simulation is a $SE(\theta) < 0.22$ with a fixed number of 12 administered items. We chose a $SE(\theta) < 0.22$ because it is comparable to a marginal reliability of .95 (Green, Bock, Humphreys, Linn, & Reckase, 1984), which results in a high standard for precise individual assessments (Bernstein & Nunnally, 1994). Regarding the fixed number of administered items, we aimed for a number that did not have a substantial impact on the precision of clinical subjects' CAT scores. We chose clinical subjects as this group is deemed the most relevant to measure anxiety (i.e., this group predominantly includes average to higher latent trait levels). To accomplish our aim, we used the criterion that at least 90% of the clinical subjects resulted in a $SE(\theta) < 0.22$. We found this number to be 12 (92%). Using this fixed number of items, we investigated whether the *SE*(θ) of the general population subjects was not compromised to a considerable extent. This was assessed by comparing the *SE*(θ) of general population subjects that did not end up with a $SE(\theta) < 0.22$ after 12 administered items, with their *SE*(θ) when no fixed number of items was applied in the stopping rule. By contrast, we also made this comparison for the number of selected items to assess the increase in administration efficiency by the fixed number of items.

As item parameters for the CAT simulation, it would be obvious to use the estimations of the complete sample. However, this would mean that we use the same data to calibrate the items and simulate the CAT, which would result in overfitting (i.e., results that are too optimistic; Hastie, Tibshirani, & Friedman, 2011). To deal with this issue, we split the clinical and general population sample randomly into half. The first half of the samples were combined to recalibrate the item bank (see "Psychometric Evaluation"

subsection); the second half of the samples were used as input for the CAT simulation. Thus, the item parameters of the complete sample (N = 2,010) could be used in a future CAT administration; the item parameters of half of the samples (n = 1,005) are used in the CAT simulation of this study. To study the similarity of the item parameters, we compared them using Pearson's correlation coefficients, and differences in means and *SD*s (complete sample parameters minus CAT simulation parameters).

Precision and Efficiency. A first demand for a CAT administration is that its outcome is both efficient (i.e., a low numbers of administered items) and precise (i.e., sufficiently free of random error). Efficiency was assessed by the mean number of selected items by the CAT simulation (and *SD*); precision was assessed by the mean $SE(\theta)$ and the percentage of respondents with a $SE(\theta) < 0.22$. In addition to these analyses, we plotted the number of selected items for each respondent as a function of the final θ estimate, along with the conditional *SE* of the Anxiety item bank. The conditional *SE* displays how precisely the item bank can measure anxiety at each level of the latent trait. It is calculated as the reciprocal square root of the sum of all item information values at each θ. All results are shown separately for the clinical and the general population sample.

Accuracy. A second demand for a CAT administration is that its outcome represents the construct which it purports to measure (i.e., free of systematic error). The θ estimates of the CAT simulation should therefore at the least be similar to those of the full item bank. We evaluated this demand by comparing the θ estimates of both tests with Pearson's correlation coefficient and Cohen's d effect size (difference between the average θ estimate divided by the pooled *SD*s). Cohen's d was calculated using the R package effsize (version 0.6.2.; Torchiano, 2016), and was evaluated using the guideline proposed by Cohen (1988): 0.2 = small effect, 0.5 = medium effect, 0.8 = large effect, a higher value meaning more systematic error between the θ estimates of the CAT simulation and those of the full item bank administration. The results are shown separately for the clinical and the general population sample.

A third demand for a CAT administration is that its outcome discriminates group membership accurately (clinical vs. healthy). The group membership assignment of the CAT simulation should therefore at the least be similar to that of the full item bank. We evaluated this demand by comparing the diagnostic accuracy of both tests (McDonald, 1999). Specifically, it was assessed how well the CAT simulation and the full item bank administration could predict the diagnostic status of a person (i.e., anxiety disorder or no disorder). For this analysis, we needed clinical subjects with an anxiety disorder and healthy persons without a disorder. Persons with an anxiety disorder were selected from the clinical sample; healthy persons (i.e., persons without current treatment for mental health issues) were selected from the general population sample. Diagnostic accuracy was assessed with the area under the curve (AUC) of the receiver operating curve, an often-used indicator for diagnostic accuracy (Rice & Harris, 2005). AUC can be interpreted as the probability that a randomly selected person with an anxiety disorder has a higher θ estimate than a randomly selected person without mental health issues (Zweig & Campbell, 1993). We used the guideline proposed by Rice and Harris (2005) to evaluate the AUC values (2005): .56 = small effect, .64 = medium effect, .71 = large effect, a higher value meaning a higher discriminative ability of the scale.

Results

Demographic Characteristics

In the clinical sample, the response rate was 31% (n = 1,032). Of the 1,032 respondents, 24 were excluded for failing to complete all 29 items. The final clinical sample therefore consisted of n = 1,008 patients (62% female; average age = 40.2 years, SD = 12.9, range: 19-76). Because the response rate of the eligible sample was only moderate, we performed a chi-square test of independence to examine whether the responders group differed from the nonresponders group. This analysis was performed for the variables gender, age, and diagnosis group (i.e., anxiety, depression, or another disorder, e.g., attention deficit disorder, somatoform disorder, personality disorder). We found no significant differences ($p < .05$) between responders and nonresponders for the variables gender and age. For the variable diagnosis group, we did find a significant difference (χ^2 [2, N = 3,296] = 11.39, $p < .05$), with somewhat less patients with a mood disorder in the responders group (44%) than in the nonresponders group (50%), somewhat more patients with an anxiety disorder in the responders group (33%) than in the nonresponders group (28%), and about an equal number of other disorders (responders group, 23%; nonresponders group, 22%). As measure of effect size, we investigated Pearson's residuals, following the suggestion of 2.00 as critical value for indicating a lack of model fit (Agresti & Kateri, 2011). It was found that only the responders group contained somewhat more patients with anxiety disorders than expected (r = 2.03).

In the general population sample, the response rate was 71% (n = 1,055). Of the 1,055 respondents, 53 respondents were excluded for showing suspicious response patterns (e.g., all responses in one category in combination with a very low response time). The final general population sample therefore consisted of n = 1,002 respondents (average age = 50.5 years, SD = 16.5, range: 19-102). The demographics of the sample were as follows: gender (male, 49%; female, 51%), age (18-39 years, 34%; 40-64 years, 44%;

65+ years, 22%), education (low, 31%; middle 40%, high 29%), ethnicity (natives, 80%; Western immigrants, 13%; non-Western immigrants, 7%), and Dutch region (north, 12%; east, 20%; south, 21%; west, 47%). Each subgroup remained within the allowed deviation of 2.5% from the Dutch population statistics in 2013.

Psychometric Properties of the Anxiety Item Bank

To begin with, the Anxiety item bank (N = 2,010) showed good descriptive statistics. Overall, the item bank showed a high internal consistency reliability (α = .98) that hardly changed when items were deleted from the item bank. Specifically, all items' scores showed a range between 1 and 5, and lacked outliers in response frequencies, mean and *SD* (see Table 1, column 3 for the item means and *SD*s). Only the item "*I felt terrified*" (EDANX33) had a minor deviation in skewness (1.16) and kurtosis (0.34). In addition, we did not find any negative or small correlation coefficients among the items. The lowest coefficient (r = 0.41) was found for item pair *I worried about other people's reactions to me* (EDANX37) and *I had twitching or trembling muscles* (EDANX44).

Next, the results from CFA and EFA indicated that the Anxiety item bank was sufficiently unidimensional. CFA showed a good fit of the unidimensional model for three out of four (scaled) fit indices: CFI = 0.97, TLI = 0.97, and SRMR = 0.04; the RMSEA indicated a moderate fit (RMSEA = 0.10). In addition, EFA showed that the first extracted factor explained 71% of the variance, which is far above the Reckase criterium of 20% (Reckase, 1979, as cited in Hambleton, 1988). Furthermore, the second extracted factor explained only 6% of the variance. The ratio of variance explained by the first to second-factor was therefore almost 12, which is 3 times higher than the required minimum of 4 (Reeve et al., 2007).

Examining the results from the residual correlation matrix and Yen's Q3 statistics, the Anxiety item bank showed sufficient LI. The residual correlation coefficients were all below the lower bound of .20 (Reeve et al., 2007), which resulted in none of the items to be marked as possibly locally dependent. With Yen's Q3 statistic, we did find two item pairs that were marked. These item pairs were *I felt fearful* (EDANX01) and *I felt frightened* (EDANX02; Q3 = .48), *I felt fearful* (EDANX01) and *I felt anxious* (EDANX05; Q3 = .42). Fortunately, removing each of these items individually from the GRM only showed a minor impact on the item parameter estimates (max 0.11 for *a*, EDANX02 and EDANX05; max 0.04 for *b*, EDANX02).

Turning to the results from the Mokken analyses, the Anxiety item bank showed monotonicity to a high degree. First, the graphs of item mean scores as a function of rest scores showed monotonicity for all items as the underlying level of the scale was higher. Second, Mokken's H was .67 for the full Anxiety item bank, which indicates a strong scale. Third, all individual items had Mokken's H values above .50 (see Table 1, column 9), which is much higher than the lower bound of .30 (Mokken, 1971).

Subsequently, the results of the OLR analyses indicated that uniform and nonuniform DIF was not present among the items of the Anxiety item bank. We confirmed this for the variables gender, age, and education level.

Finally, Table 1 (column 4 to 8) displays the GRM item parameter estimates of the Anxiety item bank. The item parameters were parametrized in the scale of the latent trait distribution of the general population sample (M = 0, SD = 1). The mean and *SD* of the clinical sample was estimated to be 1.42 and 0.70, respectively.

The item parameter estimates showed considerable variation. The discrimination parameters ranged from a = 1.34 (*I had difficulty sleeping*; EDANX49) to a = 3.59 (*I found it hard to focus on anything other than my anxiety*; EDANX40); the threshold parameters ranged from $b_1 = -0.91$ (*I felt worried*; EDANX30) to $b_4 = 4.53$ (*I had twitching or trembling muscles*; EDANX44). In addition, the p values of the $S\text{-}X^2$ statistics ranged from 0.00 to 0.36 (see Table 1, column 10). From the 29 items, 6 items had a $p < .001$ (see Table 1, column 11). These items were *I felt fearful* (EDANX01), *It scared me when I felt nervous* (EDANX03), *I was concerned about my mental health* (EDANX08), *I was easily startled* (EDANX20), *I felt something awful would happen* (EDANX27), and *I felt terrified* (EDANX33). Removing each of these items individually from the GRM only showed a minor impact on the item parameter estimates (max 0.10 for *a*, EDANX01; max 0.05 for *b*, EDANX27). We therefore concluded that the GRM fitted the Anxiety item bank sufficiently. Moreover, based on all results of the psychometric evaluation, we have chosen not to remove any of the items from the Anxiety item bank.

In Figure 1, we displayed the test information of the Anxiety item bank. The item bank is highly informative for the average and higher anxiety levels (approximately $\theta > -0.5$), and less informative for the lower anxiety levels (approximately $\theta < -0.5$). These results indicate that although we constructed a scale to measure the full latent Anxiety continuum, the item bank measures Anxiety more precisely for the average and higher anxiety levels than for the lower anxiety levels. This was to be expected as low values of the latent trait are generally related to less precise measurement in mental health constructs (Reise & Waller, 2009).

Properties of the CAT simulation

Item Parameter Estimates. The comparison between the item parameter estimates of the complete sample (N = 2,010) and the CAT simulation sample (n = 1,005) resulted in high correlation coefficients ($r_a = 1.00$, $r_{b1} = 1.00$, $r_{b2} = 1.00$, $r_{b3} = 1.00$, $r_{b4} = .99$), small differences in means

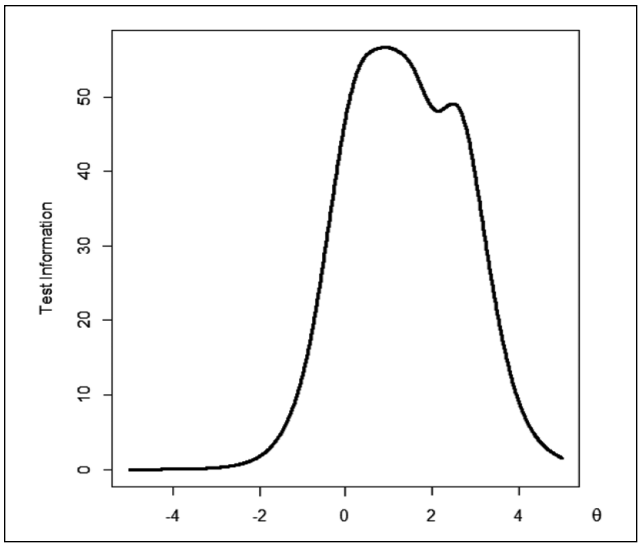

Figure 1. Test information of the Anxiety item bank.
Note. N = 2,010.

(M_a = −0.02, M_{b1} = 0.08, M_{b2} = 0.07, M_{b3} = 0.09, M_{b4} = 0.06), and small differences in SDs (SD_a = −0.02, SD_{b1} = 0.01, SD_{b2} = 0.01, SD_{b3} = 0.02, SD_{b4} = 0.03). We therefore concluded that the item parameter estimates of the CAT simulation sample are highly similar to those of the complete sample.

Efficiency and Precision. Efficient and highly precise measurement was obtained in both samples, with more gains for the clinical sample (n = 504; number of selected items, M = 8.64, SD = 1.83; mean $SE(\theta)$ = 0.22) than for the general population sample (n = 501; number of selected items, M = 9.48, SD = 2.38; $SE(\theta)$ = 0.28). This was also shown by the percentage of respondents with a $SE(\theta) < 0.22$, which was much higher in the clinical sample (92%) than in the general population sample (63%). Considering that the percentage of persons with low-anxiety values is higher in the general population, these results were to be expected (Reise & Waller, 2009).

In Figure 2, the number of selected items are displayed as a function of the final θ estimate along with the conditional SE of the Anxiety item bank. The θ estimates of the general population sample are clearly located more to the left of the scale than those of the clinical sample. At this end of the scale, the conditional SE is high. Consequently, the general population sample contained less respondents with a $SE(\theta) < 0.22$ than the clinical sample, and received more often all 12 items. By contrast, the conditional SE was at its lowest approximately between $0.00 < \theta < 2.00$. At these scale points, measurement was most efficient for the majority of respondents from both samples, with six items as the lowest number of administered items.

For the general population, we found that subjects whom did not end up with a $SE(\theta) < 0.22$ after 12 administered

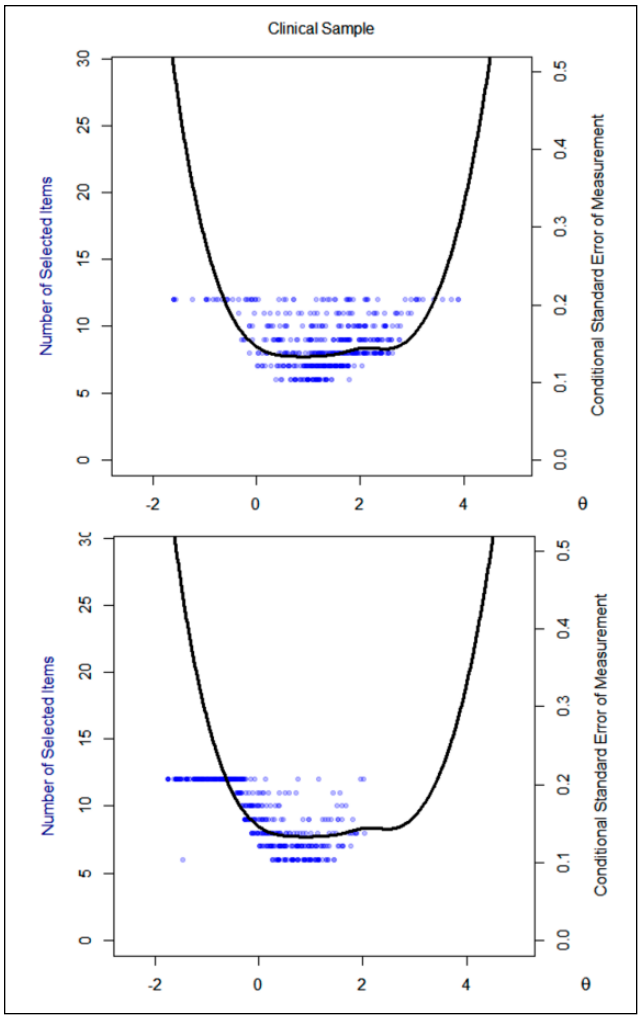

Figure 2. Number of selected items by the CAT simulation shown as a function of the final θ estimate along with the conditional standard error of measurement of the Anxiety item bank for the clinical sample and the general population sample.
Note. Stopping rule = $SE(\theta) < 0.22$ + fixed number of 12 administered items; Clinical sample, n = 504; General population sample, n = 501.

items, had an average $SE(\theta)$ = 0.39. When the CAT simulation was performed again, but without applying a fixed number of items in the stopping rule, the average $SE(\theta)$ decreased somewhat to $SE(\theta)$ = 0.35. By contrast, the mean number of selected items increased from 12.00 to 26.72. These results indicate that applying a fixed number of 12 administered items in our stopping rule did not compromise respondents' $SE(\theta)$ substantially, but did increase the administration efficiency considerably.

Accuracy. Table 2 displays Pearson's correlation coefficients and sizes of difference (Cohen's d) between the θ estimates of the CAT simulation and those of the full item bank administration. We found that the coefficients were high for both clinical and general population sample

Table 2. Pearson's Correlation Coefficient and Effect Size of the Difference (Cohen's *d*) Between the Full Anxiety Item Bank θ Estimates and the CAT Simulation θ Estimates for the Clinical Sample and the General Population Sample.

	Full θ		CAT θ			
Sample	*M*	*SD*	*M*	*SD*	*r*	*d*
Clinical	1.32	0.87	1.33	0.88	0.98	0.01
General Population	−0.11	0.96	−0.09	0.95	0.98	0.01

Note. Stopping rule = $SE(\theta) < 0.22$ + fixed number of 12 administered items; Clinical sample, $n = 504$; General population sample, $n = 501$; CAT = computerized adaptive test; *r* is Pearson's correlation coefficient; *d* is Cohen's *d*.

($r = 0.98$). Furthermore, Cohen's *d* showed a negligible effect size for both samples ($d = 0.01$). These results indicate that the θ estimates of a CAT administration may be highly similar to those of a full item bank administration.

The AUC analyses consisted of $n = 204$ patients with an anxiety disorder and $n = 449$ healthy persons. We found that the AUC value showed a large effect when the full item bank was administered (AUC = 0.92, 95% CI [0.89, 0.94]), which remained highly similar under the CAT simulation (AUC = 0.92, 95% CI [0.90, 0.95]. These results indicate that the diagnostic accuracy of a CAT administration may be highly similar to that of a full item bank administration.

Discussion

The first goal of this study was to present a psychometric evaluation of the Dutch–Flemish version of the USA PROMIS adult V1.0 item bank for Anxiety (Pilkonis et al., 2011). We used a large sample ($N = 2{,}010$) with clinical and general population subjects to demonstrate that the Anxiety item bank has desirable psychometric properties according to the PROMIS standards (Reeve et al., 2007). These properties include sufficient unidimensionality, LI, monotonicity, absence of DIF, and GRM fit. We therefore conclude that the Anxiety item bank could be used as input for a CAT administration to measure the full latent anxiety continuum. As expected, the item bank measures Anxiety more precisely for persons with average and higher anxiety levels than for persons with low-anxiety levels (Reise & Waller, 2009).

The second goal of this study was to investigate how efficient and precise a CAT version of the Anxiety item bank may be to clinical and general population subjects, and how accurate this CAT version may be compared with a full item bank administration. For this goal, we performed a post hoc CAT simulation with a stopping rule that combined a high measurement precision with a fixed number of administered items, and that was chosen with a primary focus on the measurement precision of average and higher anxiety levels. First, the simulation showed that our CAT version resulted in efficient and highly precise measurement, with more gains for the clinical sample as compared with the general population sample. For clinical practice, this may imply that measurement precision and efficiency declines somewhat as the severity of anxiety declines. This is to be expected and acceptable as the Anxiety item bank is primarily developed to measure clinical subjects. Second, the simulation showed that our CAT version was similarly accurate compared with the full item bank administration. This was shown by both θ estimates and the assignment of group membership. We therefore conclude that a CAT administration with the Anxiety item bank may not only be efficient and highly precise but also just as accurate as a full item bank administration.

In this study, we showed that the item parameter estimates of the CAT simulation sample were highly similar to those of the complete sample. This means we can assume that similar results will be obtained when our CAT version is administered with the item parameters of the complete sample. To verify this assumption, replication of the present results is necessary with a genuine CAT administration. Moreover, we need to address other accuracy aspects to validate our CAT version. These aspects include concurrent validity to ensure that our CAT version is similar to other validated anxiety instruments (McDonald, 1999), as well as longitudinal validity aspects to ensure that our CAT version could be used to asses change in respondents. Longitudinal validity aspects include measurement invariance over time (Fokkema, Smits, Kelderman, & Cuijpers, 2013; Fried et al., 2016) and responsiveness to change (de Beurs et al., 2011; Schalet et al., 2016). A measurement invariant scale means that the item bank measures the same construct at different time points; responsiveness to change means that change in θ estimates over time represent real changes in the construct (Mokkink et al., 2010).

Specific consideration should be given to the comparison of the Dutch–Flemish version of the Anxiety item bank and the original USA version. PROMIS aims to implement identical item banks and item parameters in every country to increase uniformity and enhance international comparability. This might prove difficult as the meaning of items may vary in different languages, and cultural differences may emerge across the globe regarding the valence of constructs, such as anxiety (van Widenfelt, Treffers, de Beurs, Siebelink, & Koudijs, 2005). Future research should therefore address measurement invariance between countries to assess to what extent comparisons are valid, and whether similar norms can be applied to instruments (e.g., Paz, Spritzer, Morales, & Hays, 2013; Wahl et al., 2015). Furthermore, countries should come to an international agreement about the CAT software, the CAT specifics, and the continued development of the item banks and the CAT methodology.

While awaiting these developments, our CAT version of the Dutch–Flemish PROMIS adult V1.0 item bank for Anxiety can be used in single measures. To increase the efficiency gains in these measures, the required measurement precision may be decreased, for example, to a $SE(\theta) < 0.32$, which is generally required as minimal precision for individual assessments (Bernstein & Nunnally, 1994). Using this alternative stopping rule, further simulations (the results of which are not shown herein) showed that the mean number of selected items may be decreased even further from 8.64 to 4.25 items for the clinical sample, and from 9.48 to 6.06 items for the general population sample. When the goal, however, is to assess change over time, we recommend using higher levels of measurement precision. High $SE(\theta)$ values are needed to detect true change in respondents (Brouwer, Meijer, & Zevalkink, 2013). With more precise indicators for true change, treatment providers have more useful information to assess whether to continue, change, or conclude treatment of patients. For this reason, we also recommend future researchers who are interested in change assessment to consider alternative stopping rules for the CAT administration, such as the predicted standard error reduction (Choi, Grady, et al., 2011). With this stopping rule, new items will be administered for as long as the measurement precision increases to a prespecified degree. For our CAT version, this means that the measurement precision could increase for a substantial number of respondents (see Figure 2). This stopping rule is not yet available in the Dutch CAT software, but when it does, Anxiety could be measured even more precisely.

A limitation of this study is the representativeness of the samples used. For the clinical sample, we collected data from the largest mental health institute in the Netherlands which has a broad coverage across departments over the entire country. The response rate, however, was only moderate (i.e., 31%). Furthermore, clinical subjects with an anxiety disorder were slightly overrepresented in the responders group. The difference between the responders and nonresponders group, however, was only small (i.e., approximately 5 percentage points). The effects of this selection bias will therefore likely be small. To deal with this issue in future item bank development based on clinical subjects, we recommend incorporating clinical criteria in a stratified sampling process.

In addition, the representativeness of the samples used to assess diagnostic accuracy could be somewhat improved. First, the sample size for clinical subjects with an anxiety disorder was moderately small ($n = 204$). Second, we did not have any information concerning comorbidity rates in the clinical sample. For future studies, we therefore recommend increasing the sample size for clinical subjects, and using both primary and secondary diagnostic criteria to assign group membership. In addition, the sample for healthy persons contained respondents with moderate- to high-anxiety trait levels, and may have included persons in need of treatment for their anxiety, but who either choose not to reveal being in treatment, or did not seek treatment. Ideally, to ensure a pure healthy sample, the diagnosis-free status of these respondents would be assessed with a diagnostic screener or interview, but the burden may be too high for the possible gains in classification accuracy. We therefore recommend maintaining our adopted approach in which respondents are asked whether they are currently under treatment for mental health issues. Finally, potential inclusion of anxiety disorder patients in the healthy sample likely does not bias the present results in a positive direction, but rather yields a too conservative estimate of the diagnostic accuracy of CAT.

In this study, the Dutch–Flemish version of the PROMIS adult V1.0 item bank for Anxiety was investigated. We found favorable psychometric properties, evidence of efficient and highly precise measurement applying a CAT simulation, and a similar accuracy between this CAT simulation and the full item bank administration. Similar results have been reported for the original USA version of the Anxiety item bank (Pilkonis et al., 2011; Schalet et al., 2016), the Dutch PROMIS adult V1.0 item bank for Depression (Flens et al., 2017), and other translations of the Depression item bank (e.g., Spanish, Vilagut et al., 2015; German, Jakob et al., 2015). Considering these results, the PROMIS methodology seems to fulfill its promise to measure—with an internationally applicable assessment battery—patient-reported health of adults and children more efficiently, precisely, and accurately than existing PROMs do. We therefore recommend colleagues from other countries to translate and evaluate the PROMIS item banks as input for a CAT administration.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Gerard Flens https://orcid.org/0000-0002-6683-4628

References

Agresti, A., & Kateri, M. (2011). Categorical data analysis. In *International encyclopedia of statistical science* (pp. 206-208). Berlin, Germany: Springer.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Babcock, B., & Weiss, D. J. (2013). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*, 1-18.

Baxter, A. J., Scott, K. M., Vos, T., & Whiteford, H. A. (2013). Global prevalence of anxiety disorders: A systematic review and meta-regression. *Psychological Medicine*, *43*, 897-910.

Bernstein, I. H., & Nunnally, J. C. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research*, *23*, 489-501.

Carlier, I. V. E., Meuldijk, D., van Vliet, I., van Fenema, E., van der Wee, N., & Zitman, F. (2012). Routine Outcome Monitoring and feedback on physical or mental health status: Evidence and theory. *Journal of Evaluation in Clinical Practice*, *18*, 104-110.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Hays, R. (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*, 1179-1194.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (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.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29.

Chalmers, R. P. (2015). *mirtCAT: Computerized adaptive testing with multidimensional item response theory*. Retrieved from http://CRAN.R-project.org/package=mirtCAT

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*, 1-30.

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*, 37-53.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care*, *44*, S115-S123.

Crins, M. H., Roorda, L. D., Smits, N., de Vet, H. C., Westhovens, R., Cella, D., . . . Terwee, C. B. (2015). Calibration and validation of the Dutch-Flemish PROMIS Pain Interference item bank in patients with chronic pain. *PLoS ONE*, *10*, e0134094. doi:10.1371/journal.pone.0134094

Crins, M. H., Roorda, L. D., Smits, N., Vet, H. C. W., Westhovens, R., Cella, D., . . . Terwee, C. B. (2016). Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain. *European Journal of Pain*, *20*, 284-296.

de Beurs, E., den Hollander-Gijsman, M., van Rood, Y., van der Wee, N., Giltay, E., van Noorden, M., . . . Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, *18*, 1-12.

de Graaf, R., ten Have, M., van Gool, C., & van Dorsselaer, S. (2012). Prevalence of mental disorders, and trends from 1996 to 2009: Results from the Netherlands Mental Health Survey and Incidence Study-2. *Social Psychiatry and Psychiatric Epidemiology*, *47*, 203-213.

Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the Mood and Anxiety Symptom Questionnaire. *Psychological Assessment*, *28*, 953-962.

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*, 79-105.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*, 2277-2291.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*, 520-531.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*, 1354-1367.

Fries, J., Rose, M., & Krishnan, E. (2011). The PROMIS of better outcome assessment: Responsiveness, floor and ceiling effects, and Internet administration. *Journal of Rheumatology*, *38*, 1759-1764.

Fries, J. F., Krishnan, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Research and Therapy*, *13*, R147. doi:10.1186/ar3461

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Hambleton, R. K. (1988). *Principles and selected applications of item response theory* (3rd ed.). New York, NY: American Council on Education.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Haverman, L., Grootenhuis, M. A., Raat, H., van Rossum, M. A., van Dulmen-den Broeder, E., Hoppenbrouwers, K., . . . Terwee, C. B. (2016). Dutch-Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®. *Quality of Life Research*, *25*, 761-765.

Jakob, T., Nagl, M., Gramm, L., Heyduck, K., Farin, E., & Glattacker, M. (2015). Psychometric properties of a

German translation of the PROMIS® Depression item bank. *Evaluation & the Health Professions*, *40*, 106-120.

Kim, S., Moses, T., & Yoo, H. H. (2015). Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing. *ETS Research Report Series*, *2015*, 1-19.

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., . . . Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, *62*, 278-287.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., . . . Cella, D. (2012). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research*, *21*, 739-746.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: Mouton.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*, 737-745.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Evaluation of the patient-reported outcomes information system (PROMIS®) Spanish-language physical functioning items. *Quality of Life Research*, *22*, 1819-1830.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*, 263-283.

Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. *Journal of Psychiatric Research*, *56*, 112-119.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Liu, H. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*, S22-S31.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Review of Clinical Psychology*, *5*, 27-48.

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

Revelle, W. (2013). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, *29*, 615-620.

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded responses. *Psychometrika Monograph Supplement*, *17*, 1-100.

Schalet, B. D., Pilkonis, P. A., Yu, L., Dodds, N., Johnston, K. L., Yount, S., ...Cella, D. (2016). Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. *Journal of Clinical Epidemiology*, *73*, 119-127.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*, 22-33.

Smits, N. (2015). On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: A simulation study. *Quality of Life Research*, *25*, 1635-1644.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying Computerized Adaptive Testing to the CES-D scale: A simulation study. *Psychiatry Research*, *188*, 147-155.

Smits, N., Zitman, F. G., Cuijpers, P., den Hollander-Gijsman, M. E., & Carlier, I. V. (2012). A proof of principle for using adaptive testing in Routine Outcome Monitoring: The efficiency of the Mood and Anxiety Symptoms Questionnaire-Anhedonic Depression CAT. *BMC Medical Research Methodology*, *12*, 4. doi:10.1186/1471-2288-12-4

Terwee, C. B., Roorda, L. D., de Vet, H. C. W., 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*, *23*, 1733-1741.

Torchiano, M. (2016). Package "effsize" [Software program]. Retrieved from https://cran.r-project.org/web/packages/effsize/effsize.pdf

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1-19.

van Widenfelt, B. M., Treffers, P. D., de Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review*, *8*, 135-147.

Vilagut, G., Forero, C. G., Adroher, N. D., Olariu, E., Cella, D., & Alonso, J. (2015). Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. *Journal of Psychiatric Research*, *68*, 140-150.

Voshaar, M. A. O., Peter, M., Glas, C. A., Vonkeman, H. E., Taal, E., Krishnan, E., . . . van de Laar, M. A. (2014). Calibration

of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. *PLoS ONE*, *9*, e92367. doi:10.1371/journal.pone.0092367

Wahl, I., Rutsohn, J., Cella, D., Löwe, B., Rose, M., Brähler, E., . . . Schalet, B. (2015). Does anxiety mean the same in English and German language? Evaluation of the psychometric equivalence of the PROMIS® Anxiety item bank and its German translation. *Journal of Psychosomatic Research*, *78*, 629-630.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., & Steinberg, L. (2001). Computerized adaptive testing: A primer. *Quality Life Research*, *10*, 733-734.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 109-135.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561-577.