Computerized adaptive testing in Dutch mental health care: a new tool to assess depression and anxiety

Slok-Flens, G.

Chapter 3

Development of a Computer Adaptive Test for Depression Based on the Dutch-Flemish Version of the PROMIS Item Bank

Published as:

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the health professions, 40*(1), 79-105.

3.1 Abstract

We developed a Dutch-Flemish version of the patient-reported outcomes measurement information system (PROMIS) adult v1.0 item bank for Depression as input for computerized adaptive testing (CAT). As item bank, we used the Dutch-Flemish translation of the original PROMIS item bank (28 items) and additionally translated 28 United States (US) depression items that failed to make the final US item bank. Through psychometric analysis of a combined clinical and general population sample (N = 2,010), 8 added items were removed. With the final item bank, we performed several CAT simulations to assess the efficiency of the extended (48 items) and the original item bank (28 items), using various stopping rules. Both item banks resulted in highly efficient and precise measurement of depression and showed high similarity between the CAT simulation scores and the full item bank scores. We discuss the implications of using each item bank and stopping rule for further CAT development.

Keywords: clinical assessment, computer adaptive test, depression, item response theory, PROMIS

3.2 Background

Routine outcome monitoring (ROM) is the repeated administration of questionnaires over time to monitor patients' progress towards recovery and to adapt the treatment, if indicated (Carlier et al., 2012a; de Beurs et al., 2011). In 2011, ROM has been implemented nationwide in Dutch mental health care. As ROM is used for various aims (treatment monitoring, benchmarking of institutes, and scientific research), the set of questionnaires administered to patients may become extensive which may result in diminished compliance and data loss. Consequently, more efficient measurement in (Dutch) mental health care is essential.

In 2002, the National Institutes of Health started the patient-reported outcomes measurement information system (PROMIS) initiative (Cella et al., 2007, 2010). Their main goal was to develop a new state of the art assessment system for measuring patient-reported health with highly accurate, precise, and short measures. In 2016, this ongoing initiative already brought forward a wide range of item banks (a set of questions with item parameters to measure a construct), which could be used for computerized adaptive testing (CAT). With CAT, the selection of questions is based on the answer(s) to previous questions and the assessment continues until a precise score of the measured latent construct is obtained (i.e., a score is sufficiently free of random error). For example, a patient answers the first item of a depression questionnaire with 5-point Likert-type scale items with response option 1 or 2. Consequently, the next question will be Item 5; otherwise (when response option 3 - 5 would have been chosen) the next item is Item 7. The various response categories for the follow-up question will then, in turn, lead to other items. This selection procedure based on previously given responses continues until the depression score meets the prespecified precision. By asking questions tailored to each patient, CAT can reduce administration burden with a shorter test while maintaining or even improving the precision of the test outcomes for all respondents (Fliege et al., 2005). Furthermore, CAT can select different sets of questions for patients with varying latent trait levels (θ) while the final test outcomes maintain comparability. By administering varied assessments to monitor patients over time, lack of interest in patients may also be avoided. Ultimately, these CAT benefits should decrease respondent burden, increase response rates and reduce possible bias due to selective loss of respondents (Dillman, Sinclair, & Clark, 1993).

The PROMIS initiative showed that the application of CAT results in highly efficient measurement; the PROMIS item banks show highly desirable psychometric properties (Fries, Krishnan, Rose, Lingala, & Bruce, 2011; Fries, Rose, & Krishnan, 2011; Khanna et al., 2011; Magasi et al., 2012; Pilkonis et al., 2011) and, used with CAT, result in highly accurate, precise, and short measures (Pilkonis et al., 2014). In response to these developments, the Dutch-Flemish PROMIS initiative (www.dutchflemishpromis.nl) was started in 2009 to investigate whether the PROMIS methodology could also be successfully implemented in the Netherlands. As a starting point, they translated 17 adult PROMIS item banks and 9 pediatric PROMIS item banks into Dutch-Flemish (Flemish is a variant of the Dutch language spoken in Belgium; Terwee et al., 2014). Among these item banks were the adult v1.0 item banks for mental health constructs Depression, Anxiety, and Anger (Pilkonis et al., 2011). Depression is the leading cause of disability worldwide in terms of total years lost due to disability (Marcus, Yasamy,

van Ommeren, Chisholm, & Saxena, 2012), and is the most common mental health disorder in Dutch adults (de Graaf, ten Have, van Gool, & van Dorsselaer, 2012). Therefore, the Depression item bank is an obvious choice to assess whether the PROMIS methodology could be implemented successfully in (Dutch) mental health care.

The aim of the present study was to develop a Dutch-Flemish version of the United States (US) PROMIS adult v1.0 item bank for Depression that could be used for measuring the full latent depression continuum in the Netherlands (i.e., all persons with no symptoms of depression to patients with severe depression). The US item bank comprises 28 items and is based on a selection of items from a larger item bank of 56 items (Pilkonis et al., 2011). In this 56-item bank, items were selected according to favorable psychometric qualities such as unidimensionality, local independence (LI) and monotonicity (Reeve et al., 2007). However, the selection of items for the final 28-item bank was based on the responses of a US sample. As a consequence, the selection of items may be strongly influenced by the American culture/language. Therefore, we chose to translate the original 56-item bank to investigate whether completion by Dutch respondents would result in a similar selection of items for the final item bank. For this purpose, we evaluated the psychometric properties of all 56 items. In addition, we compared the efficiency of the final item bank with the original 28-item bank by performing several post hoc CAT simulations. One of the PROMIS initiative's goals is to implement identical item banks in every country to increase uniformity and enhance comparability. By comparing the extended item bank with the original item bank, we can appraise the implications of using each item bank for further CAT development.

3.3 Method

3.3.1 Participants

For this study, data were collected in two samples: a clinical sample and a general population sample. We chose to include both samples in the item bank construction because our goal is to develop an instrument that covers the full range of possible latent depression levels in the Netherlands. Within this range, the clinical sample mostly covers moderate to high depression levels while the general population mostly covers low to moderate depression levels. We aimed to include a minimum number of 1,000 respondents per sample. A sample size of at least 1,000 is deemed sufficient for adequate item parameter estimates in the item bank calibration (Reise & Yu, 1990).

For the clinical sample, 3,296 patients were invited by the Dutch Mental Health Care provider Parnassia Psychiatric Institute to complete the item set. Patients were referred to this institute by their general practitioner for treatment of common mental disorders in ambulatory mental health care. The patient's diagnosis was assessed with the Dutch translation of the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998) administered by phone by a psychiatric nurse who was extensively trained in the interview. The MINI-plus is a standardized interview for clinical diagnosis of mental disorders following the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association,

1994). After the need for treatment was ascertained, the diagnosis was confirmed in a clinical face-to-face assessment.

According to Dutch law, use of data that are collected in the process of routine clinical practice does not require informed consent from patients. However, in accordance with the mental health-care center's policy, written informed consent was obtained.

From the general population, we needed a random sample to ensure representativeness. Respondents were invited to partake by the data collection panel Desan Research Solutions until at least 1,000 persons participated. Response rates are generally high for this panel, approximately between 60% and 80% (the total number of invitations to panel members was not registered). Respondents participated on a voluntary basis for a small financial compensation. The sample was composed to be in accordance with the Dutch general population distribution regarding five variables in 2013 (www.cbs.nl): gender (male, 49%; female, 51%), age (18-39, 34%; 40-64, 44%; 65+, 22%), education (low, 32%; middle 40%, high 28%), ethnicity (natives, 80%; western immigrants, 10%; nonwestern immigrants, 10%), and region (north, 10%; east, 21%; south, 22%; west, 47%). Deviations in each subgroup were allowed up to 2.5%.

3.3.2 Measures

The Depression item bank consisted of 28 items from the Dutch-Flemish PROMIS adult v1.0 item bank for Depression (Terwee et al., 2014), and 28 US items that did not make it to the final US PROMIS item bank (Pilkonis et al., 2011). The translation of the additional 28 US PROMIS items was performed by four researchers with ample experience in translation of self-report measures; two researchers performed a forward translation of the items, two researchers performed an independent review of these translations. Adjustments were made, until consensus was reached and the translation was approved by all four researchers. Respondents were asked for all 56 items to indicate on a Likert-type scale (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*) how frequently they experienced a wide range of depression symptoms in the past 7 days. All items reflected symptoms, problems, or negative affective states (e.g., Item 1 *I felt fearful*), a higher score meaning more severe depression.

3.3.3 Psychometric Evaluation

We performed a psychometric evaluation of the 56-item bank on the combined patient and general population sample, following the guidelines proposed by Reeve et al. (2007). First, we evaluated several descriptive statistics to assess the performance of the individual depression items and the full Depression item bank. Individual items were evaluated with response frequencies and range, mean, standard deviation (*SD*), skewness, and kurtosis. Furthermore, we explored the interitem correlation matrix, the item-scale correlations, and the drop in coefficient α for each item when removed from the item bank. In addition, the full item bank was evaluated with the sum score range, mean, SD, skewness, kurtosis, and the reliability coefficient for internal consistency.

Second, we evaluated the main item response theory (IRT) assumptions of unidimensionality, LI, and monotonicity to assess whether the Depression item bank is fit to

scale respondents and items on a common latent trait. Item banks are considered *unidimensional* when a person's item response results from the person's trait level that the item measures and not from other factors. However, mental health constructs are generally complex and rarely strictly unidimensional. For IRT applications, it is therefore assessed whether the degree of unidimensionality in item banks is sufficient (Reise, Morizot, & Hays, 2007). Unidimensionality was evaluated with exploratory factor analyses (EFA) using the R package psych (Version 1.5.4; Revelle, 2013), and with confirmatory factor analyses (CFA) using the R package lavaan (Version 0.5-18; Rosseel, 2012), both conducted on the polychoric correlation matrix of the items (Bollen, 1989). With EFA, unidimensionality is deemed sufficient when the first factor accounts for at least 20% of the variance (Reckase, 1979, as cited in Hambleton, 1988), and the ratio of explained variance in the first and second factor is higher than 4 (Reeve et al., 2007). With CFA, unidimensionality of the Depression item bank is deemed sufficient when the comparative fit index (CFI) > 0.95, the Tucker-Lewis index (TLI) > 0.95, the root-mean-square error of approximation (RMSEA) < 0.06, and the average absolute residual correlations < 0.10 (Reeve et al., 2007).

The second IRT assumption we evaluated is *LI*. Item pairs are locally independent when, controlling for the trait level, item responses show no association. LI in the depression items was evaluated by inspecting the residual correlation matrix that resulted from the single-factor CFA. Residual correlations higher than .20 were considered as possibly locally dependent (Reeve et al., 2007). Further investigation of LI was done with Yen's Q3 statistic (Yen, 1993), in which the residual item scores under Samejima's graded response model (GRM; Samejima, 1969), fitted with R package mirt (Version 1.10; Chalmers, 2012), are correlated among items. As suggested by Smits, Cuijpers, and van Straten (2011), model fit was evaluated with Cohen's (1988) rules of thumb to interpret effect size; Q3 values between 0.24 and 0.36 imply moderate deviations, Q3 values above 0.37 imply large deviations.

The third IRT assumption we evaluated is *monotonicity*. Items show monotonicity when the probability of selecting an item response that suggests a better health status on a scale increases as the underlying level of health status on that scale is higher. Monotonicity in the depression items was evaluated by examining graphs of item mean scores conditional on rest scores (total raw score minus the item score), using the R package mokken (Version 2.7.7; van den Ark, 2007). This analysis additionally results in scalability coefficients for the full scale and the individual items. A scale or item has low quality when the scalability coefficient is between 0.30 and 0.40, moderate quality when the scalability coefficient is between 0.40 and 0.50, and high quality when the scalability coefficient is above 0.50 (Mokken, 1971).

Subsequently, we evaluated *differential item functioning* (DIF; Embretson & Reise, 2000) to assess whether persons from different groups have equal probabilities of selecting item response categories. An item shows DIF when the probability of responding in different response categories differs across independent groups, controlling for the trait level influencing a person's item response. We explored DIF for gender (men, women), age (18-39, 40-64, 65+), and education level (low, medium, and high). DIF among the depression items was evaluated with ordinal logistic regression (OLR; Crane, Gibbons, Jolley, & van Belle, 2006), using the R package lordif (Version 0.2-2; Choi, Gibbons, & Crane, 2011). Effect size was evaluated by

means of change in McFadden's pseudo R^2, following the suggestion of a critical value of 0.02 (Choi et al., 2011) for rejecting the hypothesis of no DIF.

Finally, we calibrated the extended item bank with Samejima's GRM (Samejima, 1969), using the R package mirt (Version 1.10; Chalmers, 2012). We fitted the GRM with multiple group estimation (McDonald, 1999; Smits, 2016) for which we used the combined clinical and general population sample and specified population as grouping factor with constraints on equal discrimination and threshold parameters. The latent trait was scaled to a mean of 0 and a *SD* of 1 for the general population. In addition, we performed a calibration on the original 28 items of the PROMIS adult v1.0 item bank for Depression to compare efficiency results with the extended item bank (see "CAT Simulations" subsection). Note that from here on all items from the original item bank are mentioned as "original item" and all additional items from the extended item bank are mentioned as "added item".

The calibrations of the extended and original item bank under the GRM were evaluated by examining item fit and item properties. First, item fit was evaluated with the $S\text{-}X^2$ statistic (Orlando & Thissen, 2000, 2003), which compares 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). Second, item properties were evaluated by examining *a* (discrimination) and *b* (threshold) parameter estimates. The discrimination parameter represents the extent to which persons with similar scores on the latent trait can be differentiated by the item. The four threshold parameters *b* (the number of threshold parameters for an item is equal to the number of response categories minus one) represent the θ locations on which a person is expected to choose from a lower to a higher item response. In addition, we compared the item parameter estimates of the first 28 items between the extended and the original item bank, using differences in means and SD's (extended minus original), and Pearson's correlations.

3.3.4 CAT Simulations

To assess the efficiency of the extended and the original item bank, we performed an individual post hoc CAT simulation with each item bank, using the R package mirtCAT (Version 0.5; Chalmers, 2015). A CAT simulation is not an actual CAT administration, but selects the item responses and evaluates them as if they had been collected adaptively. We split the clinical and general population samples randomly into half; the first half of both samples was used for estimating the item parameters, the second half for simulating CAT. This method will prevent overfitting (Hastie, Tibshirani, & Friedman, 2001), which would have resulted in outcomes that are too optimistic. Note that we estimated the item parameters again to perform this analysis. Thus, the item parameters of the full clinical and general population sample are to be used as input for a future CAT, the item parameters of half of the clinical and general population sample are used in this study as input for simulating CAT.

We chose to perform the primary CAT simulations on the clinical sample because clinical subjects were deemed the most relevant group to measure depression. In addition, we also performed CAT simulations with each item bank using the general population sample and briefly mention some main results. It could be expected that the efficiency gains are higher for

the clinical sample compared to the general population sample because the information value of items is generally lower for respondents with low values of the latent trait (low levels of depression; Reise & Waller, 2009).

The CAT simulations started with the item that had the highest item information value at the average value of the latent trait (θ = 0; Embretson & Reise, 2000; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). Consequently, the CAT simulations started with the original item Emotional Distress – Depression item bank, Item 36 (EDDEP36) *I felt unhappy* for both the extended and the original item bank (note that we use the original US item coding; www.assessmentcenter.net). The depression latent trait scores (θ) were then estimated with the Bayesian method maximum a posteriori (MAP; Embretson & Reise, 2000), and a standard error (*SE*) was calculated. The CAT simulation stopped selecting new items when the patient's θ reached a prespecified value of the *SE*. Otherwise, new items were selected using the highest item information at the provisional θ estimate until the prespecified value of the *SE* was obtained or when all items were selected without obtaining the *SE*. We evaluated several stopping rules: $SE(\theta) < 0.1$, $SE(\theta) < 0.2$, $SE(\theta) < 0.3$, and $SE(\theta) < 0.4$. For each stopping rule, several statistics were recorded individually for both the extended and the original item bank to assess the efficiency of CAT: (a) the mean and *SD* of the number of selected items, (b) the percentage of all patients for whom all items had to be selected, and (c) the mean *SE* of the final θ estimate for all patients. In addition, we investigated the efficiency of CAT under each stopping rule by plotting the number of selected items for each patient with the test information of each item bank. Test information displays how precisely an item bank can measure a latent trait, given the location of the person's estimate. It is calculated as the sum of all item information values at any relevant θ level.

3.3.5 Comparing full-scale data with CAT data

Through CAT simulations, we could assess the similarity between patients' estimated CAT θ scores and patients' estimated full item bank θ scores. For this analysis, we used the patients from the (CAT simulation) clinical sample (n = 504). First, similarity between the depression scores was assessed with Pearson's correlation. Second, we assessed the effect size between both depression scores using Cohen's d (with pooled *SD*'s), which was evaluated using the guideline proposed by Cohen (1988): 0.2 = small effect, 0.5 = medium effect, and 0.8 = large effect. We performed these analyses for the original and the extended item bank under all stopping rules.

The CAT simulations also enabled us to assess whether depressed persons systematically differed in θ estimates from persons without a diagnosis, as a minimal requirement for predictive validity. For this analysis, we compared scores of persons with a mood disorder (n = 161) to the scores from persons without a diagnosis (n = 449). Cohen's d (with pooled *SD*'s) was assessed for the original and the extended item bank under all stopping rules, including the full-scale estimates (no stopping rule).

3.4 Results

3.4.1 Demographic characteristics

From the 3,296 invited patients, 1,032 completed the questionnaire (response rate = 31.3%). We did not find differences between responders and nonresponders for the variables gender and age. Among the 1,032 respondents, 24 patients were excluded from the analyses because they did not complete all 56 items. Therefore, the final clinical sample consisted of n = 1,008 patients (61.7% female). The mean age of the patients was 40.2 years (*SD* = 12.9, range 19–76). Patients' diagnoses (*DSM-IV*) were classified as follows: 44% had a mood disorder, 33% had an anxiety disorder, and 23% had another disorder (e.g., attention deficit disorder, somatoform disorder, personality disorder, etc.).

From the 1,055 respondents of the general population, 53 persons were excluded because they showed suspicious response patterns (e.g., all responses in one category). Therefore, the final general population sample consisted of n = 1,002 persons from the Dutch population. The mean age of the general population sample was 50.5 years (*SD* = 16.5, range 19–102). Regarding demographics, the sample was composed as follows: gender (male, 49%; female, 51%), age (18-39, 34%; 40-64, 44%; 65+, 22%), education (low, 31%; middle 40%; high 29%), ethnicity (natives, 80%; western immigrants, 13%; nonwestern immigrants, 7%), and region (north, 12%; east, 20%; south, 21%; west, 47%).

3.4.2 Psychometric evaluation

For the psychometric evaluation of the data, the clinical sample and the general population sample were combined (56 items; N = 2,010). The extended item bank data did not show outliers in response frequencies of the depression items, mean, *SD*, range, skewness, and kurtosis. Furthermore, the data showed a high internal consistency reliability (Cronbach's α = .99). However, the added item EDDEP11 *I ate more than usual* showed a very small negative interitem correlation (r = -0.02) with the added item EDDEP49 *I lost weight without trying*. This negative correlation is also implied by the content of the items, as the item *I ate more than usual* is implicitly about gaining weight.

All CFA fit indices resulted in a good fit (CFI = 0.99; TLI = 0.99; average absolute residual correlations = 0.04), except for the RMSEA, which resulted in a moderate fit (RMSEA = 0.09). With EFA, the proportion of variance explained by the first factor was 68% which is above the Reckase criterion of 20% (Reckase, 1979, as cited in Hambleton, 1988). In addition, the ratio of variance explained by the first and second factor was 17, which was also higher than the minimal requirement of 4 (Reeve et al., 2007). We concluded that the extended item bank sufficiently met the assumption of unidimensionality.

Of all item pairs, 8 added item pairs were considered possibly locally dependent as their residual correlations were above .20 (Reeve et al., 2007). Further investigation of these items with Yen's Q3 statistic showed 3 item pairs with high deviations (item pairs EDDEP32 *I wished I were dead and away from it all* – EDDEP33 *I thought about suicide*, EDDEP32 *I wished I were dead and away from it all* – EDDEP40 *I felt that others would be better off if I were dead*, and EDDEP33 *I thought about suicide* – EDDEP40 *I felt that others would be better off if I*

were dead), 3 item pairs with moderate deviations (item pairs EDDEP11 *I ate more than usual* – EDDEP15 *I disliked the way my body looked*, EDDEP49 *I lost weight without trying* – EDDEP53 *I had little desire to eat*, and EDDEP16 *I felt like crying* – EDDEP34 *I had crying spells*), and 2 item pairs with no deviations (item pairs EDDEP11 *I ate more than usual* – EDDEP49 *I lost weight without trying*, and EDDEP11 *I ate more than usual* – EDDEP53 *I had little desire to eat*). Items with residual correlations > .20, high deviations in Yen's Q3 statistic, and other poor psychometric properties were removed from the item bank.

The graphs of item mean scores conditional on rest scores showed monotonicity for all 56 depression items. In addition, the scalability coefficient of the Depression item bank was high (.64), and the scalability coefficient for all depression items was above the lower bound of .30 (Mokken, 1971). We concluded that the extended item bank sufficiently met the assumption of monotonicity.

The 56 depression items showed no DIF for age and education level. For gender, the added items EDDEP16 *I felt like crying* and EDDEP34 *I had crying spells* were flagged for DIF. Change in McFadden's R^2 was .03 for both items, which was above the threshold of .02 (Choi et al., 2011).

Based on the statistical results, we chose to remove the added items EDDEP11 *I ate more than usual*, EDDEP49 *I lost weight without trying*, EDDEP16 *I felt like crying*, and EDDEP34 *I had crying spells*. First, item EDDEP11 *I ate more than usual* and EDDEP49 *I lost weight without trying* for having a small negative correlation with each other. Both are symptoms of depression, but cannot occur at the same time in a single person. Therefore, the item response for one of these items could result in bias because it is not clear which item can be seen as a depression symptom in a person. Second, we removed item EDDEP16 *I felt like crying* and EDDEP34 *I had crying spells* for having DIF on gender. Based on content, we additionally chose to remove the added items EDDEP53 *I had little desire to eat* and EDDEP55 *I felt like I needed help for my depression*. First, EDDEP53 *I had little desire to eat* because just as items EDDEP11 *I ate more than usual* and EDDEP49 *I lost weight without trying* both the confirmation and the rejection of this item can be seen as a depression symptom in different persons. Second, EDDEP55 *I felt like I needed help for my depression* because this item is not appropriate for healthy respondents. After removing these items, we reevaluated all psychometric qualities of the extended 50-item bank and found that they had all improved slightly.

In the calibration of the remaining items, we found five $S\text{-}X^2$ *p*-values below .001 for the extended 50-item bank (original items EDDEP42 and EDDEP46; added items EDDEP32, EDDEP38 and EDDEP40) and seven $S\text{-}X^2$ *p*-values below .001 for the original 28-item bank (original items EDDEP09, EDDEP21, EDDEP27, EDDEP39, EDDEP42, EDDEP44 and EDDEP54). Based on content and other psychometric properties, we chose to remove the added items EDDEP32 *I wished I were dead and away from it all* and EDDEP40 *I felt that others would be better off if I were dead*; both items showed a high degree of local dependency with item EDDEP33 *I thought about suicide*. After the 48-item bank was recalibrated, we did not find other items that needed to be removed. In addition, the correlation between the estimated latent trait scores (θ) under the full item banks (extended item bank, 48 items; original item

bank, 28 items) and the sum of raw scores under the full item banks was high for both the original and the extended item bank (r = .99). We concluded that the GRM fitted the extended 48-item bank and the original 28-item bank sufficiently.

Table 3.1 Discrimination and threshold parameter estimates for the extended and original PROMIS item bank for Depression.

		Extended item bank					Original item bank				
Item code	Item	Item parameter estimates					Item parameter estimates				
		a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4
EDDEP04	I felt worthless	2.718	-0.109	0.606	1.412	2.528	2.940	-0.030	0.680	1.493	2.618
EDDEP05	I felt that I had nothing to look forward to	2.638	-0.063	0.615	1.423	2.500	2.747	0.007	0.687	1.508	2.611
EDDEP06	I felt helpless	2.781	0.038	0.772	1.530	2.657	2.818	0.104	0.847	1.625	2.793
EDDEP07	I withdrew from other people	2.400	-0.170	0.573	1.408	2.610	2.336	-0.113	0.646	1.506	2.765
EDDEP09	I felt that nothing could cheer me up	3.446	0.091	0.746	1.467	2.321	3.493	0.156	0.818	1.561	2.456
EDDEP14	I felt that I was not as good as other people	2.672	0.006	0.641	1.329	2.257	2.746	0.074	0.714	1.418	2.373
EDDEP17	I felt sad	2.693	-0.504	0.218	1.195	2.309	2.700	-0.446	0.285	1.279	2.437
EDDEP19	I felt that I wanted to give up on everything	2.700	0.143	0.810	1.604	2.599	2.672	0.204	0.885	1.707	2.746
EDDEP21	I felt that I was to blame for things	2.666	0.246	0.934	1.614	2.493	2.689	0.312	1.010	1.711	2.626
EDDEP22	I felt like a failure	3.243	0.360	0.904	1.494	2.254	3.498	0.430	0.978	1.581	2.364
EDDEP23	I had trouble feeling close to people	2.048	0.040	0.865	1.711	2.837	1.963	0.095	0.946	1.826	3.011
EDDEP26	I felt disappointed in myself	3.209	-0.141	0.497	1.249	2.196	3.272	-0.076	0.564	1.335	2.325
EDDEP27	I felt that I was not needed	2.627	0.121	0.805	1.545	2.522	2.644	0.186	0.877	1.641	2.661
EDDEP28	I felt lonely	2.702	-0.120	0.556	1.314	2.214	2.723	-0.055	0.628	1.401	2.334
EDDEP29	I felt depressed	3.450	0.114	0.713	1.304	2.126	3.388	0.174	0.784	1.400	2.265
EDDEP30	I had trouble making decisions	2.429	-0.289	0.533	1.381	2.436	2.241	-0.253	0.597	1.489	2.621
EDDEP31	I felt discouraged about the future	3.232	-0.113	0.512	1.198	2.100	3.316	-0.049	0.577	1.280	2.219
EDDEP35	I found that things in my life were overwhelming	2.564	0.216	0.919	1.704	2.660	2.422	0.272	0.999	1.825	2.843
EDDEP36	I felt unhappy	3.946	-0.103	0.557	1.186	2.068	4.111	-0.034	0.626	1.269	2.192
EDDEP39	I felt I had no reason for living	2.578	0.842	1.349	1.965	2.732	2.581	0.913	1.437	2.079	2.870
EDDEP41	I felt hopeless	3.935	0.482	1.016	1.690	2.520	3.904	0.540	1.090	1.803	2.691
EDDEP42	I felt ignored by people	2.276	0.383	1.156	2.051	3.185	2.143	0.441	1.247	2.199	3.403
EDDEP44	I felt upset for no reason	2.717	0.377	0.980	1.670	2.821	2.488	0.432	1.061	1.795	3.042
EDDEP45	I felt that nothing was interesting	3.283	0.266	0.885	1.554	2.496	3.075	0.322	0.960	1.667	2.682
EDDEP46	I felt pessimistic	3.057	0.047	0.671	1.377	2.350	2.978	0.105	0.736	1.471	2.502
EDDEP48	I felt that my life was empty	3.431	0.204	0.769	1.342	2.196	3.469	0.268	0.842	1.435	2.333
EDDEP50	I felt guilty	2.773	0.195	0.807	1.502	2.429	2.683	0.255	0.880	1.605	2.585
EDDEP54	I felt emotionally exhausted	3.496	-0.017	0.556	1.085	1.952	3.124	0.032	0.617	1.171	2.107
EDDEP12	I had mood swings	2.592	0.133	0.715	1.453	2.418					
EDDEP43	I felt slowed down	2.688	0.167	0.737	1.484	2.502					
EDDEP10	I was critical of myself for my mistakes	2.219	-0.027	0.590	1.344	2.256					
EDDEP56	I had trouble enjoying things that I used to enjoy	3.137	0.041	0.560	1.142	2.080					
EDDEP13	I felt that other people did not understand me	2.845	-0.052	0.601	1.328	2.294					
EDDEP51	I lost interest in my appearance	1.694	0.297	1.147	2.089	3.171					
EDDEP03	I felt that I had no energy	2.458	-0.229	0.391	1.113	2.091					
EDDEP08	I felt that everything I did was an effort	2.919	-0.255	0.413	1.142	2.173					
EDDEP15	I disliked the way my body looked	1.503	-0.031	0.795	1.635	2.586					
EDDEP18	I got tired more easily than usual	2.246	-0.029	0.508	1.184	2.252					
EDDEP24	I felt like being alone	1.863	-0.123	0.503	1.397	2.641					
EDDEP01	I reacted slowly to things that were said or done	2.447	0.327	1.051	1.899	2.767					
EDDEP20	My thinking was slower than usual	2.365	0.179	0.885	1.676	2.679					
EDDEP33	I thought about suicide	1.868	1.342	1.887	2.658	3.547					
EDDEP38	I felt unloved	2.572	0.624	1.220	1.802	2.651					
EDDEP47	I had trouble keeping my mind on what I was doing	2.447	-0.051	0.575	1.356	2.619					
EDDEP52	I had trouble thinking clearly	2.720	0.128	0.755	1.540	2.582					
EDDEP37	I was unable to do many of my usual activities	2.146	0.629	1.259	1.973	3.086					
EDDEP02	I felt lonely even when I was with other people	3.229	0.507	0.937	1.590	2.478					
EDDEP25	I had bad dreams that upset me	1.746	0.735	1.407	2.150	3.207					

In Table 3.1, the final item parameter estimates of the extended 48-item bank and the original 28-item bank are displayed (N = 2,010; clinical sample, n = 1,008 and general population sample, n = 1,002). The item parameter estimates of both the extended 48-item bank and the original 28-item bank showed considerable variation. For the extended 48-item bank, the item parameter estimates ranged from a = 1.503 (added item EDDEP15 *I disliked the way my body looked*) to a = 3.946 (original item EDDEP36 *I felt unhappy*), and from b_1 = -0.504 (original item EDDEP17 *I felt sad*) to b_4 = 3.547 (added item EDDEP33 *I thought about suicide*). For the original 28-item bank, the item parameter estimates ranged from a = 1.963 (EDDEP23 *I had trouble feeling close to people*) to a = 4.111 (EDDEP36 *I felt unhappy*), and from b_1 = -0.446 (EDDEP17 *I felt sad*) to b_4 = 3.403 (EDDEP42 *I felt ignored by people*). In addition, the comparison between the matching 28 items of the extended and original item bank showed high Pearson's correlations (r_a = .97, r_{b1} = 1.00, r_{b2} = .99, r_{b3} = .97, r_{b4} = .96), small differences in means (m_a = 0.02, m_{b1} = 0.17, m_{b2} = 0.19, m_{b3} = 0.25, m_{b4} = 0.36), and small differences in *SD*s (SD_a = -0.04, SD_{b1} = 0.00, SD_{b2} = -0.02, SD_{b3} = -0.07, SD_{b4} = -0.12).

3.4.3 Efficiency of CAT using different stopping rules

In Table 3.2, the CAT simulation outcomes for the clinical sample are displayed for the extended and original item bank under each stopping rule (n = 504). Evidently, both the mean number of selected items and the number of patients for whom the full item banks were selected declined, as the stopping rule was less strict.

Table 3.2 Patients' CAT simulation statistics for the extended and original PROMIS item bank for Depression under several stopping rules.

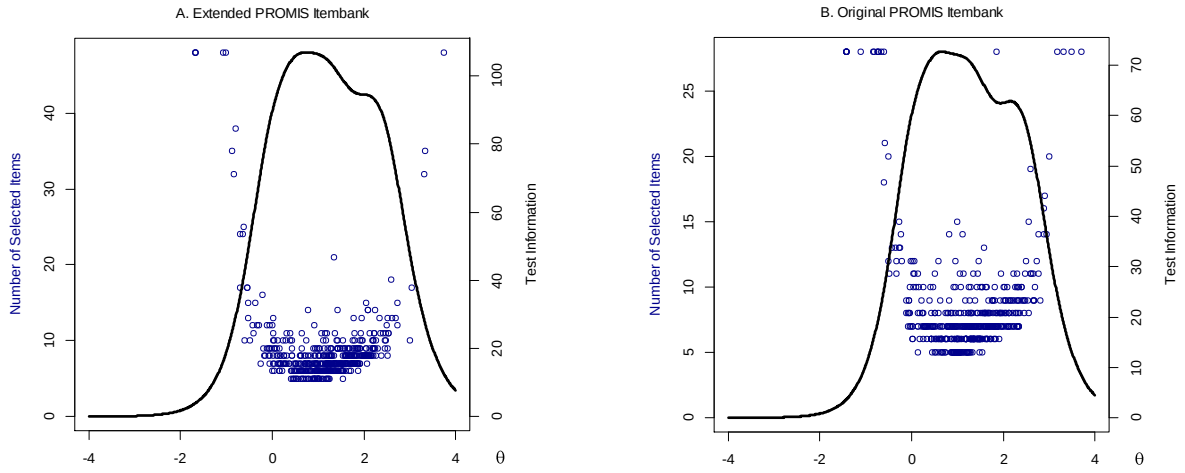
	Extended item bank				Original item bank			
Stopping rule	Number of items				Number of items			
	M	*SD*	% All	Mean *SE*(θ)	*M*	*SD*	% All	Mean *SE*(θ)
SE(θ) <0.1	44.29	5.07	56.2	.11	28.00	0.00	100.0	.13
SE(θ) <0.2	8.69	5.68	1.2	.20	8.40	4.45	3.6	.20
SE(θ) <0.3	3.48	4.04	0.6	.28	3.40	3.33	1.4	.28
SE(θ) <0.4	2.09	3.83	0.6	.35	2.03	2.76	1.0	.35

Apart from stopping rule $SE(\theta) < 0.1$, the extended and original item bank show highly similar results. Apparently, stopping $SE(\theta) < 0.1$ is too strict for both item banks as the simulations selected all items for a high percentage of patients (Table 3.2, column 4 and 8). This is especially the case with the original item bank (100% full item bank selections) due to its relative low number of items (28 in the original item bank to 48 in the extended item bank). From stopping rule $SE(\theta) < 0.2$, however, the mean number of selected items dropped substantially for both item banks, following a similar pattern (Table 3.2, column 2 and 6). Under stopping rule $SE(\theta) < 0.2$, the mean number of selected items is around 8.54, and then dropped even further to 3.44 under stopping rule $SE(\theta) < 0.3$, and 2.06 under stopping rule $SE(\theta) < 0.4$. These stopping rules also result in a much smaller percentage of patients for whom all items

were selected (below 4%). Overall, the efficiency of the original item bank is slightly higher. This result is an effect of the large difference in the number of items in each item bank. As a consequence, the mean number of selected items from the extended item bank is somewhat inflated by the group of patients for whom (almost) all items were selected.

As example, Figure 3.1 shows the test information along with the number of selected items under stopping rule $SE(\theta)$ <0.2 for both the extended (1A) and the original (1B) item bank. Evidently, test information is higher for most θ values in the extended item bank due to the larger number of items. However, the shape of the test information curve is similar for both item banks, meaning that test information is high for $-0.5 < \theta < 3$ and low for $\theta < -0.5$ (very low depression score) or $\theta > 3.0$ (very high depression score). Obviously, the number of selected items is linked to the test information, because large number of items were selected for patients with θ estimates at the end of the scales (low-test information). In contrast, only 5 or 6 items were selected for most patients with θ estimates in the middle of the scale (high-test information). This pattern was shown for all stopping rules for both item banks, naturally with a decline in number of selected items as the stopping rule was less strict. Under stopping rule $SE(\theta) < 0.4$, for example, only 1 or 2 items were selected for patients with θ estimates that showed high-test information.

Figure 3.1 Number of selected items shown as a function of the final θ estimate under stopping rule $SE(\theta) < 0.2$ for the extended and original PROMIS item bank.



Finally, the CAT simulation outcomes for the general population showed, as expected, less efficiency gains compared to the clinical sample. Naturally, most respondents from the general population had θ estimates at the lower end of the depression scale (very low depression scores), which indicates very low-test information. Consequently, the mean number of selected items increased. For example, under stopping rule $SE(\theta) < 0.2$, the mean number of selected items was 19 with the extended item bank and 14 with the original item bank.

3.4.4 Comparing full-scale data with CAT data

In Table 3.3, Pearson's correlations and sizes of difference (Cohen's *d*) between patients' CAT simulation θ estimates and patients' full item bank θ estimates are displayed for the extended and original item bank under each stopping rule (n = 504). Note that the results regarding the mean and *SD* of both item banks cannot be compared directly. Because the datasets are different (i.e., the number of items), the metric of the scales is also slightly different. As a result, the extended and the original item bank show a small difference in mean and *SD* of the θ estimates (extended item bank: full-scale θ, M = 1.15 and SD = 0.79; original item bank: full-scale θ, M = 1.21 and SD = 0.83).

Table 3.3 Pearson's correlations and sizes of difference (Cohen's d) between patients' CAT simulation θ estimates and patients' full item bank θ estimates for the extended and original PROMIS item bank for Depression under several stopping rules.

	Extended Item Bank				Original Item Bank			
	CAT θ				CAT θ			
Stopping Rule	*M*	*SD*	*r*	*d*	*M*	*SD*	*r*	*d*
$SE(\theta) < 0.1$	1.15	.79	1.00	.00	1.21	.83	1.00	.00
$SE(\theta) < 0.2$	1.14	.78	.96	.01	1.19	.81	.97	.02
$SE(\theta) < 0.3$	1.11	.78	.92	.05	1.14	.81	.94	.08
$SE(\theta) < 0.4$	1.03	.78	.87	.14	1.08	.78	.89	.14

Evidently, Pearson's correlations declined and sizes of difference increased as the stopping rule was less strict. Again, the extended and original item bank showed highly similar results. Pearson's correlations were high under all stopping rules, ranging from 1.00 under stopping rule $SE(\theta) < 0.1$ to 0.87 (extended item bank) and 0.89 (original item bank) under stopping rule $SE(\theta) < 0.4$ (Table 3.3, column 4 and 8). In addition, Cohen's *d* values indicated a negligible to a very small effect under all stopping rules, ranging for both item banks from 0 under the stopping rule $SE(\theta) < 0.1$ to 0.14 under stopping rule $SE(\theta) < 0.4$ (Table 3.3, column 5 and 9). Specifically, patients' mean CAT simulation θ estimates under stopping rule $SE(\theta) < 0.1$ were equal to patients' mean full item bank θ estimates, and declined as the stopping rule was less strict. For clinical practice this would imply that less strict stopping rules yield slightly lower depression scores with CAT.

Table 3.4 presents the sizes of difference (Cohen's *d*) between the θ estimates of persons with a mood disorder and the θ estimates of persons without a diagnosis for the extended and original PROMIS item bank under several stopping rules. Cohen's *d* was large under each stopping rule and nearly identical for the extended and original item bank, ranging from 1.41 (extended item bank) and 1.45 (original item bank) under the full item bank (no stopping rule) to 1.22 for both item banks under stopping rule $SE(\theta) < 0.4$. The results indicate that depressed patients have a much higher θ estimate on the depression scale than persons without a diagnosis, and this difference declines somewhat when the stopping rule is less strict.

Table 3.4 Sizes of difference (Cohen's d) between the θ estimates of persons with a mood disorder and persons without a diagnoses for the extended and original PROMIS item bank for Depression under several stopping rules.

	Extended item bank					Original item bank				
	Mood disorder		No diagnosis			Mood disorder		No diagnosis		
Stopping rule	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
None: θ	.94	.72	-.24	.88	1.41	.99	.75	-.22	.86	1.45
$SE(\theta) < 0.1$	.94	.72	-.24	.88	1.41	.94	.72	-.22	.86	1.41
$SE(\theta) < 0.2$	.91	.69	-.22	.88	1.35	.91	.69	-.21	.87	1.36
$SE(\theta) < 0.3$	.89	.68	-.17	.85	1.32	.89	.68	-.16	.84	1.32
$SE(\theta) < 0.4$	.82	.71	-.13	.80	1.22	.82	.71	-.12	.80	1.22

3.5 Discussion

In this study, we evaluated the Dutch-Flemish version of the US PROMIS adult v1.0 item bank for Depression with data from a sample of patients with mental health problems and a sample from the Dutch general population. We started with a 56-item bank that was also used in the US validation study (Pilkonis et al., 2011). In the US, the validation of the Depression item bank resulted in 28 items (original item bank). Although all 28 items were retained in our study, we found a total of 48 items with desirable psychometric qualities (extended item bank). These psychometric qualities included sufficient unidimensionality, LI and monotonicity, and absence of DIF. Furthermore, the 48-item bank showed a sufficient fit with the GRM (Samejima, 1969).

We compared the original and extended item bank using a post hoc CAT simulation and found that the efficiency of both item banks for patients was highly similar, with a slight superiority for the original item bank. Therefore, based on efficiency, the original bank could also be used for CAT implementation. Using the smaller 28-item bank has the additional benefit of enhanced international comparability between the Dutch-Flemish and the US item banks for the assessment of depression. To investigate comparability further, future research should address factorial invariance and DIF between countries.

CAT methodology is not only aimed at improving efficiency, but also at varied assessments for patients with differing θ estimates. Within the treatment process, this CAT characteristic is also advantageous with repeated administrations over time. By administering varied assessments to monitor patients' health progress, diminished attentiveness may be avoided. This benefit might be more clearly visible in the extended item bank than the original item bank due to the larger number of items. For future research, we therefore recommend the US PROMIS group to assess whether the original 28-item US PROMIS adult v1.0 item bank for Depression could be extended with newly validated items.

Using CAT to assess respondents, it is common to adopt stopping rule $SE(\theta) < 0.3$ (e.g., Becker et al., 2008; Gibbons et al., 2014). This stopping rule is comparable to a marginal reliability of .90 (Green, Bock, Humphreys, Linn, & Reckase, 1984), which is generally

required for minimal reliability for individual assessments (Bernstein & Nunnally, 1994, p. 265). Our findings suggest that stopping rule $SE(\theta) < 0.3$ would be a sound choice for using CAT with the original or the extended Depression item bank. Under this stopping rule, the mean number of selected items was very low (extended item bank, $m = 3.48$; original item bank, $m = 3.40$), patients' CAT simulation θ estimates showed a sufficient similarity with patients' full item bank θ estimates, and patients with a depression diagnosis differed substantially in θ estimates from persons without a diagnosis. However, our ultimate goal is to use CAT in ROM to monitor patients' progress over time. To assess significant change, high levels of individual test precision are required. Significant change can be expressed with the IRT-based *Z*-test (Brouwer, Meijer, & Zevalkink, 2013) using pretest and posttest data. The pre-post difference needed to deem a patient as significantly changed is dependent of the *SE* of measurement. With a lower *SE* of pretest and posttest, we will be better able to detect true change. It may therefore be more suited to use a stopping rule requiring more precision such as $SE(\theta) < 0.2$. Under this stopping rule, the mean number of selected items is still very acceptable (extended item bank, $m = 8.69$; original item bank, $m = 8.40$), the similarity between patients' CAT simulation θ estimates and patients' full item bank θ estimates is substantial for both item banks, and depressed patients differed substantially in θ estimates from persons without a diagnosis.

When choosing a stopping rule, researchers should also take into account the maximum number of items the CAT software should administer to increase the efficiency for each individual. This can be done by setting a fixed number of maximum items, or by incorporating (state of the art) stopping rules which take into account whether additional items will increase the precision or change the estimated latent trait value of the assessment (e.g., predicted SE reduction [PSER], Choi, Grady, & Dodd, 2011; change in θ, Babcock & Weiss, 2013). Using one of these methods is especially useful for persons with very high or very low depression levels. For such persons, test information is low which could result in the administration of all items in the item bank without ever meeting the *SE* stopping rule. Limiting the maximum number of items to be administered should therefore not result in an unacceptable diminishment of precision of the test result. Consequently, the slight inferiority of the extended item bank in this study should diminish because the mean number of selected items is no longer affected by individuals for whom all or most of the 48 items were selected.

After choosing the item bank and stopping rule, the Dutch-Flemish version of the PROMIS adult v1.0 item bank for Depression could be used in clinical practice for single measure purposes to assess the level of depressive symptomatology. For utilizing CAT specifically in diagnostic prediction, future research needs to further address predictive validity using patients´ diagnoses. For utilizing CAT in ROM, future research needs to address measurement invariance over time (Fokkema, Smits, Kelderman, & Cuijpers, 2013) and whether responsiveness to change of CAT θ estimates is equal to full item bank θ estimates or to responsiveness of legacy measures (de Beurs et al., 2012).

A possible limitation of the present study is that the results regarding the efficiency of the item banks were assessed with CAT simulations and not with real CAT administrations. Although another study has shown that the outcomes of CAT simulations and real CAT administrations can be very similar (Kocalevent et al., 2009), replications of these results are

necessary. For example, the CAT simulations results were based on item parameters from a smaller sample (n = 1,004) and therefore could differ somewhat from real CAT administration results that are based on item parameters from a larger sample (N = 2,008). Furthermore, the correlations and the sizes of difference between patients' CAT simulation θ estimates and patients' full item bank θ estimates could respectively be inflated and deflated as the data derive from the same assessment. An independent administration with both the full item bank and the CAT in the same subjects could provide useful information about the utility of CAT simulations to assess CAT efficiency gains.

Another factor that should be taken into account when using CAT is the influence of shrinkage in the θ estimates of the Bayesian estimation method MAP (Embretson & Reise, 2000). Shrinkage basically means that the use of a prior normal distribution pulls θ estimations towards the mean, especially with early θ estimations. As a consequence, θ could be somewhat over- or underestimated for patients with a low number of selected items. This effect might explain the slightly diminishing mean in θ estimates as the stopping rules were less strict (Table 3.3). A solution to deal with the influence of shrinkage in Bayesian θ estimation is by setting a minimum number of items the CAT should administer or by using a different estimation method (e.g., maximum likelihood; Smits, 2016).

In this study, we showed that the PROMIS methodology results in efficient measurement of depression in Dutch patients. The Dutch-Flemish PROMIS item banks (extended and original) show desirable psychometric qualities, and applied as a CAT, could result in short and precise measurement. These favorable results were also found in other countries using different translations of the PROMIS item banks (e.g., German, Jakob et al., 2015; Spanish, Vilagut et al., 2015). We therefore encourage researchers in other countries to investigate whether the PROMIS methodology is efficient and valid for assessment of depression in their clinical and general population.