Quality of Life Research (2019) 28:211–220
https://doi.org/10.1007/s11136-018-1995-0

Validation of two PROMIS item banks for measuring social participation in the Dutch general population

C. B. Terwee[1] · M. H. P. Crins[2] · M. Boers[1,3] · H. C. W. de Vet[1] · L. D. Roorda[2]

Accepted: 4 September 2018 / Published online: 10 September 2018

Abstract

Background The Patient-Reported Outcomes Measurement Information System (PROMIS) item banks 'Ability to Participate in Social Roles and Activities' (35 items) and 'Satisfaction with Social Roles and Activities' (44 items) were developed to measure (satisfaction with) participation more efficiently and precisely than current instruments, by using Computerized Adaptive Testing (CAT). We validated these item banks in a Dutch general population.

Methods Participants in an internet panel completed both item banks. Unidimensionality, local dependence, monotonicity, Graded Response Model item fit, Differential Item Functioning (DIF) for age, gender, education, region, ethnicity, and language (Dutch compared to US Social Supplement), and reliability were assessed.

Results A representative Dutch sample of 1002 people participated. We found for the Ability to Participate and Satisfaction with Participation item banks, respectively, sufficient unidimensionality (CFI: 0.971, 0.960; TLI: 0.970, 0.958; RMSEA: 0.108, 0.108), no local dependence, sufficient monotonicity (H: 0.75, 0.73), good item fit (2 out of 35 items, 1 out of 44 items with S-X^2 p-value < 0.001). No DIF was found. We found a reliability of at least 0.90 with simulated CATs in 86% and 94% of the participants with on average 4.7 (range 2–12) and 4.3 (range 3–12) items, respectively.

Discussion The PROMIS participation item banks showed sufficient psychometric properties in a general Dutch population and can be used as CAT. PROMIS CATs allow reliable and valid measurement of participation in an efficient and user-friendly way with limited administration time.

Keywords Participation · Validation · PROMIS · IRT

Introduction

Participation in social roles and activities is a major determinant of many favorable health and quality of life outcomes, and a key dimension of successful aging [1, 2]. The *International Classification of Functioning, Disability and Health (ICF)* distinguishes participation from activity limitations (or physical function) and defines participation as "an individual's involvement in life situations in relation to health conditions, body functions and structure, activities, and contextual factors" [3]. Social participation declines as a result of 'normal' aging [4] and increasing morbidity. It is important to monitor social participation in populations and individual patients to develop and evaluate interventions that can improve social participation, for example in elderly and patients with chronic diseases [5–7], and to measure participation as an outcome in clinical trials [8, 9].

Many Patient-Reported Outcome Measures (PROMs) are available to measure participation, both generic and disease-specific [8, 10–14], but their use is associated with a number of challenges. The instruments vary in content and operationalization of the concept of participation [10–12] and in their measurement properties [11, 15–17]. Scores of different instruments are incomparable because the scales are ordinal and even if scores are expressed on a scale from 0 to 100, one cannot assume that a score of 40 points on one instrument corresponds to a score of 40

✉ C. B. Terwee
cb.terwee@vumc.nl

[1] Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands

[2] Amsterdam Rehabilitation Research Center | Reade, Amsterdam, The Netherlands

[3] Amsterdam Rheumatology and Immunology Center, VU University Medical Center, Amsterdam, The Netherlands

points on another instrument. Furthermore, many instruments are considered to be too long for use in daily clinical practice.

The Patient-Reported Outcomes Measurement Information System (PROMIS) initiative developed two universally applicable item banks to measure participation, defined as performing one's social roles and activities: the PROMIS item bank 'Ability to Participate in Social Roles and Activities' and the PROMIS item bank 'Satisfaction with Social Roles and Activities' [18, 19]. The item banks were developed based on modern psychometric methods (Item Response Theory (IRT)) to overcome the above-mentioned challenges.

PROMIS item banks have several advantages compared to traditional PROMs. One of the main advantages is that PROMIS item banks can be administered as Computerized Adaptive Test (CAT). In CAT, after the first item, the selection of subsequent items is determined by the person's responses to the previous items. With CAT, persons need to complete on average only 3–7 items to get a reliable score, compared to 20–30 items with a traditional questionnaire [20, 21]. This makes PROMIS CATs highly suitable for use in studies, alongside other instruments, as well as in daily clinical practice. Fixed short forms of subsets of 4–8 items are also available for applications where a computer is not available. Short forms of the Ability to Participate item bank are included in PROMIS Profile instruments [22], and in preference measures [23, 24]. Another advantage of PROMIS is that the instruments are applicable across (disease) populations. The interest in universally applicable instruments is rising, given the increasing number of people with multiple chronic diseases [25–28].

Good psychometric properties of the PROMIS participation item banks were found in US and Spanish general populations. Both language versions were considered unidimensional (Comparative Fit Index (CFI) 0.97 for the Ability item bank in English and Spanish and 0.96 and 0.94 for the Satisfaction item bank in English and Spanish, respectively), all items fitted the IRT model and item parameters were considered similar (i.e., no Differential Item Functioning (DIF)) across gender, age, and education [19]. Evidence for internal consistency, test–retest reliability, construct validity, and responsiveness was found for the short forms in patients with rheumatoid arthritis, osteoarthritis, fibromyalgia, systemic lupus erythematosus, systemic sclerosis, idiopathic pulmonary fibrosis, and patients undergoing cervical spine surgery [29–34].

The PROMIS participation item banks were recently translated into Dutch-Flemish [35]. The aim of this study was to validate these two item banks in a general Dutch population.

Methods

Participants

We used an existing internet panel of the general Dutch population polled by a certified company (Desan Research Solutions). We deemed a sample of at least 1000 people sufficient for item parameter estimation. The sample should be representative of the Dutch general population (maximum of 2.5% deviation) with respect to distribution of age (18–40; 40–65; > 65), gender, education (low, middle, high), region (north, east, south, west), and ethnicity (native, first, and second generation western immigrant, first and second generation non-western immigrant), based on data from Statistics Netherlands in 2016 [36].

Measures

The item bank Ability to Participate in Social Roles and Activities assesses the perceived ability to perform one's usual social roles and activities. All 35 items are worded in terms of perceived limitations, e.g., "I have trouble doing my regular daily work around the house" and scored on a 5-point Likert response scale (never, rarely, sometimes, usually, always). Responses are reverse-coded so that higher scores represent better ability. The item bank Satisfaction with Social Roles and Activities assess satisfaction with performing one's usual social roles and activities, e.g., "I am satisfied with my ability to participate in family activities." All 44 items are scored on a 5-point Likert response scale (not at all, a little bit, somewhat, quite a bit, very much) with higher scores representing more satisfaction. There is no time frame in any of the items.

Procedure

Participants completed all 35 items of the Dutch-Flemish V2.0 PROMIS item bank Ability to Participate in Social Roles and Activities and all 44 items of the Dutch-Flemish V2.0 PROMIS item bank Satisfaction with Social Roles and Activities through an online survey. In addition, participants completed general questions about age, gender, education, region, and ethnicity.

Statistical analysis

We conducted psychometric analyses in accordance with the PROMIS analysis plan [37]. An IRT model requires that three assumptions are met: unidimensionality, local independence, and monotonicity.

First, we examined unidimensionality by Confirmatory Factor Analyses (CFA) on the polychoric correlation matrix with Weighted Least Squares with Mean and Variance adjustment (WLSMV) estimation, using the R package LAVAAN (version 0.5-23.1097) [38]. For unidimensionality, all items must load on a single factor. The CFI, Tucker Lewis Index (TLI), Root Means Square Error of Approximation (RMSEA) and Standardized Root Mean Residual (SRMR) evaluated model fit. We report scaled fit indices, which are considered more exact than unscaled indices [39, 40]. In addition, we performed an Exploratory Factor Analysis (EFA) on the polychoric correlation matrix with WLSMV estimation procedures using the R package Psych (version 1.7.5) [41]. Following the PROMIS analysis plan and recommendations from Hu and Bentler [42], we considered sufficient evidence for unidimensionality if CFI > 0.95, TLI > 0.95, RMSEA < 0.06, SRMR < 0.08, the first factor in EFA accounted for at least 20% of the variability, and the ratio of the variance explained by the first to the second factor was greater than four [37, 43].

Second, we evaluated local independence. After controlling for the dominant factor, there should be no significant covariance among item responses. We examined the residual correlation matrix resulting from the single factor CFA mentioned above, and considered residual correlations greater than 0.20 indicators of possible local dependence [37].

Third, we assessed monotonicity. The probability of endorsing a higher item response category should increase (or at least not decrease) with increasing levels of the underlying construct [37]. We evaluated monotonicity by fitting a non-parametric IRT model, with Mokken scaling, using the R-package Mokken (version 2.8.4) [44, 45]. This model yields non-parametric IRT response curve estimates, showing the probabilities of endorsing response categories that can be visually inspected to evaluate monotonicity. We evaluated the fit of the model by calculating the scalability coefficient H per item and for the total scale. We considered monotonicity acceptable if the scalability coefficients of the items were at least 0.30, and the scalability coefficient for the total scale was at least 0.50 [45].

To study IRT model fit, we fitted a logistic Graded Response Model (GRM) to the data using the R-package Mirt (version 3.3.2) [46]. A GRM models two kind of item parameters, item slopes and item thresholds. The item slope refers to the discriminative ability of the item, with higher slope values indicating better ability to discriminate between adjoining values on the construct. Item thresholds refer to item difficulty, and locate the items along the measured trait (i.e., the construct of interest) [47]. For items with five response categories, four item thresholds are estimated. To assess the fit of the GRM model, we used a generalization of Orlando and Thissen's S-X^2 for polytomous data [48]. This statistic compares observed and expected response frequencies under the estimated IRT model, and quantifies the differences between them. The criterion for good fit of an item is a S-X^2 p-value greater than 0.001 [49].

We used DIF analyses to examine measurement invariance, i.e., whether people from different groups (e.g., males vs females) with the same level of the trait (participation) have different probabilities of giving a certain response to an item [50]. We evaluated DIF for age (median split: ≤ 53 years, > 53 years), gender (male, female), education (low, middle, high), region (north, east, south, west), ethnicity (native, first and second generation western immigrant, first and second generation non-western immigrant), and language (English vs Dutch). For this latter aim, we compared our sample to the US PROMIS 1 Social Supplement, obtained from the HealthMeasures Dataverse repository [51], which was used to develop these item banks [18]. We selected only the participants from this Supplement who were recruited from the US general population (Polimetrix sample, $n = 1008$). From this group, we used 429 people with complete data for the Ability to Participate in Social Roles and Activities item bank and 424 people with complete data for the Satisfaction with Social Roles and Activities item bank for the DIF analyses. We evaluated DIF by a series of ordinal logistic regression models, using the R package Lordif (version 0.3-3) [52], which models the probability of giving a certain response to an item as a function of the trait, a (dichotomous or ordinal) group variable, and the interaction between the trait and the group variable. We used a McFadden's pseudo R^2 change of 2% between the models as a criterion for DIF [52]. Uniform DIF exists when the magnitude of the DIF is consistent across the entire range of the trait. Non-uniform DIF exists when the magnitude or direction of DIF differs across the trait.

Finally, we evaluated reliability. Reliability within IRT is conceptualized as "information." Information (I) is inversely related to the standard error (SE) of the estimated construct or trait level (called theta, θ), as indicated by the formula:

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

The SE can differ across theta [47, 53]. The theta is estimated based on the GRM model and scaled with a mean of 0 and a SD of 1 and an effective range of −4 to 4. An SE of 0.316 corresponds to a reliability of 0.90 (SE 0.548 corresponds to a reliability of 0.70). For each person, we calculated four theta scores: one based on all items of the item bank, one based on the standard 8-item short form (version 8a), and two based on different CAT simulations. In the first simulated CAT, we used the standard PROMIS CAT stopping rules. The standard CAT stopped when a SE of 3 on the T-score metric was reached (comparable to a reliability

slightly higher than 0.90) or a maximum of 12 items was administered (the recommended minimum by PROMIS is 4 items, but this could not be defined in catR so we used no minimum). In the second simulated CAT, we administered a fixed number of 8 items to compare the reliability of the CAT with the short form. In all simulations, the starting item was the item with the highest information value for the average level of participation in the population ($\theta = 0$), according to PROMIS practice. We used the R-package catR (version 3.12) for the CAT simulations [54]. We used maximum likelihood (ML) for estimating thetas to prevent biased scores in people with extreme responses [55]. ML, however, is not able to estimate θ for response patterns that exclusively comprise extreme responses. Therefore, we set the possible scale boundaries in the CAT simulation to − 4 to 4, whereby people who score 1 or 5 on all CAT items get a theta score of − 4 or 4.

We transformed theta scores into *T*-scores as recommended by PROMIS according to the formula (θ*10) + 50. A *T*-score of 50 represents the average score of the study population, with a standard deviation of 10. We plotted the SE across *T*-scores for the entire item banks, for the standard 8-item short forms (version 8a), and for the two different CAT simulations [54]. We plotted the distribution of *T*-scores in our population to show the reliability of the item bank in relation to the distribution of scores in the population.

Results

A sample of 1002 people from the panel completed the questionnaire (mean age 51 (SD 17), 52% female) between July and November 2016. All participants had complete data. The demographic characteristics of the participants are summarized and compared to the Dutch general population in 2016 in Table 1. All differences were less than the 2.5% agreed upon.

Ability to Participate in Social Roles and Activities

The three assumptions for fitting an IRT model were considered to be met. The scaled CFA fit indices were CFI: 0.97, TLI: 0.97, RMSEA: 0.11, and SRMR: 0.04. In EFA, the eigenvalue of the first factor was 27.3; the eigenvalue of the second factor was 0.92 (ratio 29.7). These results were considered showing enough evidence for unidimensionality. No item pairs were flagged for local dependence. The Mokken scalability coefficients of the items were ≥ 0.30 (range 0.64–0.79) and the scalability coefficient of the full item bank was 0.75, supporting monotonicity.

Four out of 35 items had a poor item fit in the GRM model, with S-X^2 *p*-value of less than 0.001 (RP1 "I have trouble doing my regular daily work around the house," SRPPER09_CaPS "I have trouble doing everything for work that I want to do (include work at home)," SRPPER17r1 "I feel limited in the amount of time I have for my family," and SRPPER28r1 "I have to limit my regular activities with friends"). The item slope parameters ranged from 2.4 to 4.8, with mean of 3.9. The item with lowest slope (worst discriminative ability) was SRPPER43r1 ("I have trouble keeping in touch with others"), and the item with the highest slope (best discriminative ability) was SRPPER08_CaPS ("I have trouble doing all of the family activities that are really important to me"). The item threshold parameters ranged from − 2.5 to 0.6. No items were flagged for DIF for age, gender, education, region, ethnicity, or language.

Based on the GRM model with ML estimation, a theta could not be estimated for 77 (7.7%) of the participants because they had extreme scores. The item with the highest information at $\theta = 0$ (average of the population) was SRPPER08_CaPS "I have trouble doing all of the family activities that are really important to me." This item was used as a starting item in the CAT simulations.

Figure 1 shows the standard error across *T*-scores for the full item bank, the short form, and the two simulated CATs. The full item bank (35 items) had a reliability of > 0.90 for 92% of the participants. The short form (8 items) had a reliability of > 0.90 for 85% of the participants. Using the standard CAT with SE = 3 and max 12 items, a reliability of > 0.90 was obtained for 86% of the participants with an average of 4.7 (range 2–12) items. Using a fixed 8-item CAT, a reliability of > 0.90 was obtained for 82% of the participants.

The mean *T*-score of the study sample (based on the full item bank, obtained from https://www.assessmentcenter.net/ac_scoringservice, using US item parameters) was 50.6 (SD 9.5), range of 20.5–69.3.

Satisfaction with social roles and activities

The three assumptions for fitting an IRT model were considered to be met. The scaled CFA fit indices were CFI: 0.96, TLI: 0.96, RMSEA: 0.11, and SRMR: 0.05. In EFA, the eigenvalue of the first factor was 33.0; the eigenvalue of the second factor was 1.44 (ratio 22.9). These results were considered showing enough evidence for unidimensionality. No item pairs were flagged for local dependence. The Mokken scalability coefficients of the items were ≥ 0.30 (range 0.66–0.77), and the scalability coefficient of the full item bank was 0.73, supporting monotonicity.

Two out of 44 items had a poor item fit in the GRM model, with a S-X^2 *p*-value of less than 0.001 (SRPSAT24r1 "I am satisfied with my ability to work (include work at home)," and SRPSAT51r1 "I am satisfied with my ability to run errands"). The item slope parameters ranged from 2.3 to 4.2, with mean of 3.5. The item with lowest slope (worst

Table 1 Characteristics of the participants

	Dutch general population study sample	Dutch general population 2016[a]	Difference between study sample and population in 2016	US sample[b]
	N (%)	%	%	*N* (%)
Number of patients	1002	13,562,539		1008
Age (years)				
18–39	316 (31.5)	33.7	–2.2	
40–65	457 (45.6)	43.6	2.0	
> 65	229 (22.9)	22.7	0.2	
Mean (*SD*) *in years*	51 (17)			56 (15)
Gender				
Male	477 (47.6)	49.2	–1.6	336 (40.8)
Female	525 (52.4)	50.8	1.6	487 (59.2)
Education[c]				
Low	294 (29.3)	30.2	–0.9	3 (0.3)
Middle	427 (42.6)	40.2	2.4	171 (17.0)
High	281 (28.0)	29.6	–1.6	823 (81.6)
Region[d]				
North	102 (10.2)	10.2	0.0	
East	199 (19.9)	20.8	–0.9	
South	201 (20.1)	21.6	–1.5	
West	497 (49.6)	47.4	2.2	
Ethnicity				
Native	774 (77.2)	78.6	–1.4	White: 74.4%
First and second generation western immigrant	127 (12.6)	10.3	2.4	
First and second generation non-western immigrant	101 (10.1)	11.2	–1.1	

[a]Based on data from statistics Netherlands (http://www.cbs.nl)

[b]Of the total sample of 1008, we used 429 people with complete data for the Ability to Participate in Social Roles and Activities item bank and 424 people with complete data for the Satisfaction with Social Roles and Activities item bank. Number of missing values: age: 6, gender: 185, education: 185, ethnicity: 188

[c]Low = primary school, lower levels of secondary school (in Dutch: VMBO or lower), lower vocational education; middle = higher levels of secondary school (in Dutch: HAVO/VWO), middle vocational education; high = at least first year of bachelor degree

[d]Three missing values in the Dutch sample

discriminative ability) was SRPSAT48r1 "I am satisfied with my ability to do things for fun at home (like reading, listening to music, etc.)," and the item with the highest slope (best discriminative ability) was SRPSAT29_CaPS "I am satisfied with my ability to engage in activities with friends." The item threshold parameters ranged from −2.1 to 1.6. None of the items were flagged for DIF for age, gender, education, region, ethnicity, or language.

Based on the GRM model with ML estimation, a theta could not be estimated for 28 (2.8%) of the participants because they had extreme scores. The item with the highest information at $\theta = 0$ (average of the population was SRPSAT29_CaPS) "I am satisfied with my ability to engage in activities with friends." This item was used as a starting item in the CAT simulations.

Figure 2 shows the standard error across *T*-scores for the full item bank, the short form, and the two simulated CATs. The full item bank (44 items) had a reliability of > 0.90 for 97% of the participants. The short form (8 items) had a reliability of > 0.90 for 95% of the participants. Using the standard CAT with SE = 3 and max 12 items, a reliability of > 0.90 was obtained for 94% of the participants with an average of 4.3 (range 3–12) items. Using a fixed 8-item CAT, a reliability of > 0.90 was obtained for 93% of the participants.

The mean *T*-score of the study sample (based on the full item bank, obtained from https://www.assessmentcenter.net/

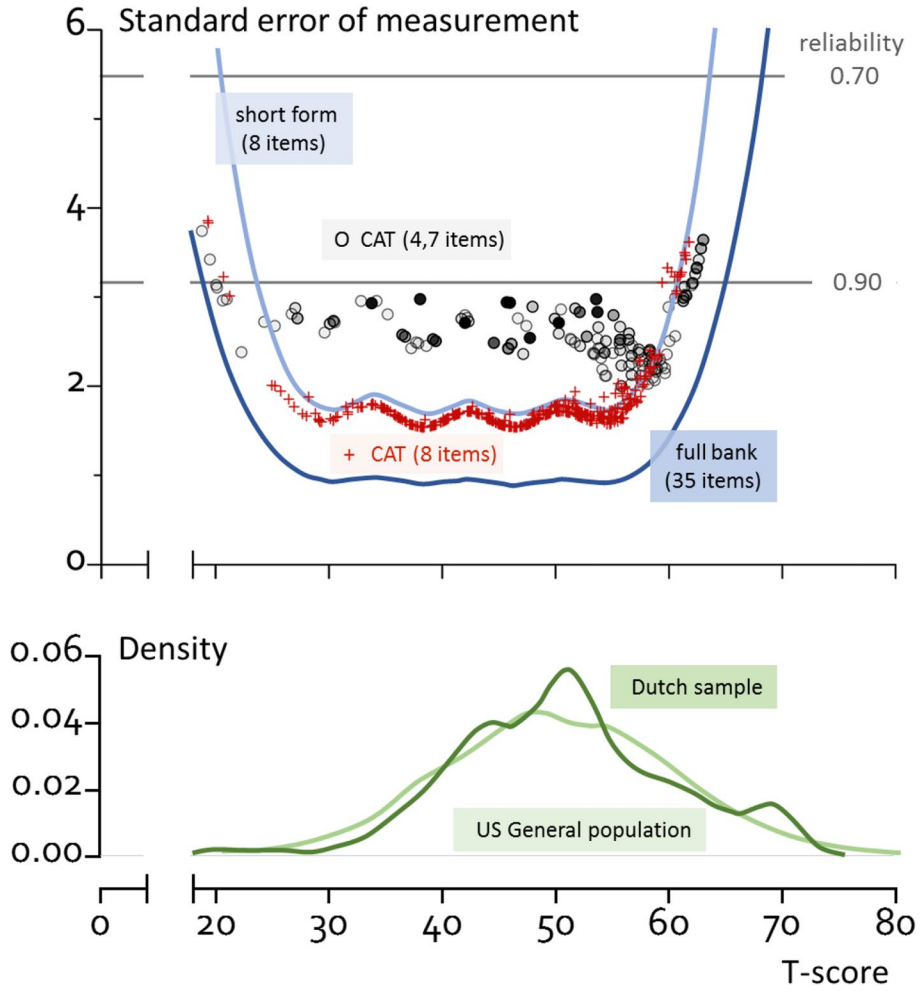

Fig. 1 Reliability of the PROMIS V2.0 item bank Ability to Participate in Social Roles and Activities when using different applications (full item bank, short form and simulated CATs (the open circle represent the standard CAT (shading represents many of the same scores) and the plus symbols represent the fixed 8-item CAT)) and distribution of *T*-scores (based on full item bank) in the population. For $n=77$, theta could not be estimated and CAT theta scores were set to −4 (*T*-score 10, $n=7$) or 4 (*T*-score 90, $n=70$) in both simulated CATs (these persons are not shown in the plot)

ac_scoringservice, using US item parameters) was 47.5 (SD 8.3), range of 20.4–70.8.

Discussion

We validated the Dutch-Flemish PROMIS item banks V2.0 Ability to Participate in Social Roles and Activities and V2.0 Satisfaction with Social Roles and Activities in a Dutch general population. It comprises (after Spanish) the second foreign language validation and the first validation in the Netherlands. We found sufficient unidimensionality, no local dependence, sufficient monotonicity, good IRT model fit, and a high reliability across a wide range of the construct for both item banks. We found no evidence for DIF due to age, gender, education, region, ethnicity, or language.

For both item banks, we found CFI and TLI values higher than the minimum criteria of 0.95, and comparable to those found in the first validation study, performed in the US and Spanish population [19]. However, the RMSEA was higher than the maximum criterion of <0.06 (0.11 for both item banks). The RMSEA was not reported for the US and Spanish population. A high RMSEA has been reported for many other PROMIS item banks [56–60]. It has been suggested that traditional cutoffs and standards for CFA fit statistics are not suitable to establish unidimensionality of item banks measuring health concepts [61] and that the RMSEA is sensitive to model complexity (number of estimated parameters) and skewed data distributions [61], the latter being the case in health concepts. Reise et al. have found the RMSEA statistic to be problematic for assessing unidimensionality of health concepts, and considered the SRMR more promising to determine whether a scale is 'unidimensional enough' [62].

Four items of the Ability to Participate in Social Roles and Activities item bank and two items of the Satisfaction with Social Roles and Activities item bank showed poor model fit. These items have low to moderate item slopes and low to moderate item information as compared to the other items (data not shown), which implicates that they do not have a high probability to be selected in a CAT. Therefore, these items will likely not cause any problems in CAT

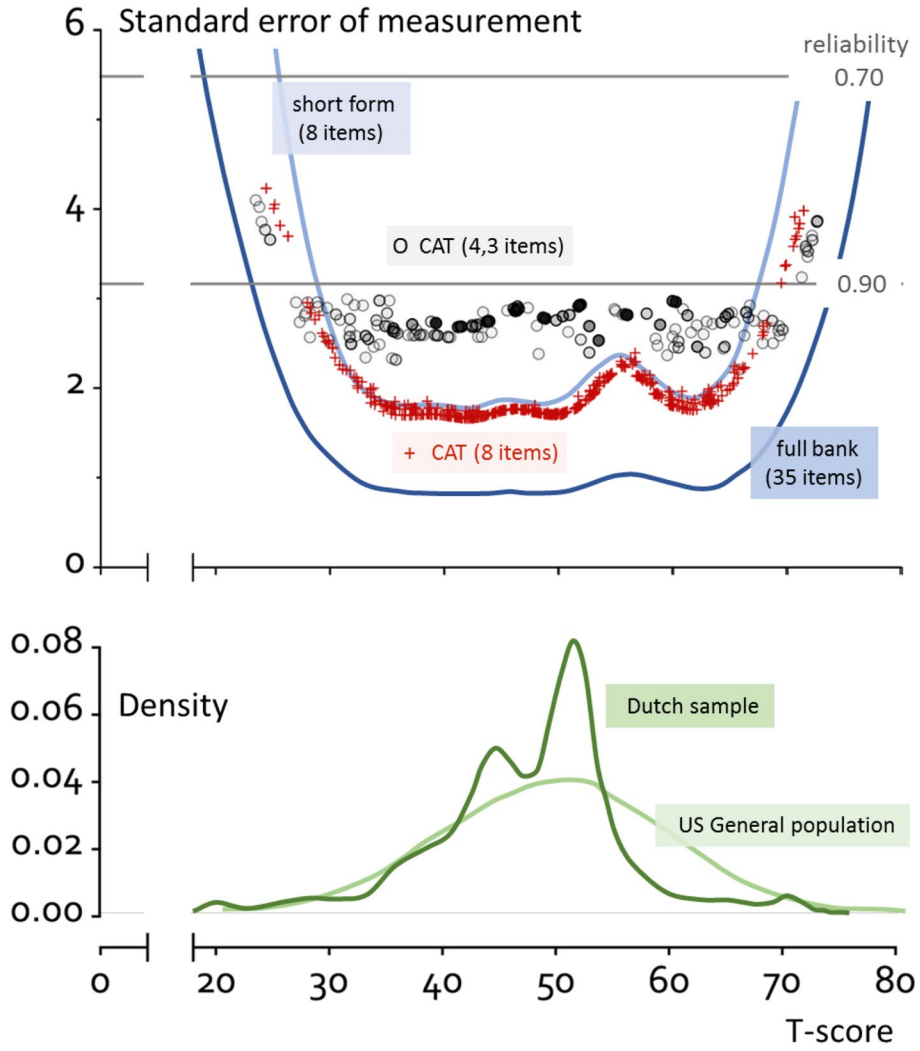

Fig. 2 Reliability of the PROMIS V2.0 item bank Satisfaction with Participation in Social Roles and Activities when using different applications (full item bank, short form and CATs (the open circle symbols represent the standard CAT (shading represents many of the same scores) and the plus symbols represent the fixed 8-item CAT)) and distribution of T-scores (based on full item bank) in the population. For $n = 28$ theta could not be estimated and CAT theta scores were set to -4 (T-score 10, $n = 12$) or 4 (T-score 90, $n = 16$) in both simulated CATs (these persons are not shown in the plot)

administrations. We prefer to keep them in the item bank at the moment because they may have higher information value in specific populations, such as patient groups. This has to be examined in future studies.

We found high reliability of the full item banks, short forms, and simulated CATs. A reliability of > 0.90 (which has been considered a minimum requirement for use of PROMs in individual patients [63]) was found in 85% and 97% of the participants for the 8-item short form, in 82% and 95% of the participants for the 8-item CAT, and in 86% and 94% of the participants for the standard CAT (mean 4.7 and 4.3 items) for the Ability and Satisfaction bank, respectively. These results show two things: First, the short forms and fixed 8-item CATs perform similar in this population. This was to be expected because the short forms include the items that best cover the full construct in the general population. Our results therefore confirm the validity of the short forms. One should keep in mind, however, that in clinical populations, with lower scores on average, short forms may perform less well because their content is then suboptimal. Second, standard CATs perform as well as short forms in this population, but use on average only four to five items instead of eight. In fact, 71% and 54% of the participants needed to complete only two to three items of the Ability and Satisfaction item bank, respectively.

This study has limitations. Although the agreed maximum of 2.5% deviation was met, the study sample was not perfectly representative of the Dutch general population: middle aged people, middle educated people, Western immigrants, females, and participants from the Western part of the Netherlands were slightly overrepresented. This will most likely not have affected the model parameters, since no DIF was found for age, gender, education, region, and ethnicity. The US sample was on average older, contained more women and more highly educated people than the Dutch sample. This will most likely not have affected the DIF for language results, since no DIF was found for age, gender, and education.

With regard to the IRT analyses, it is not clear what the best estimation method is to estimate theta scores. We

used maximum likelihood (ML) estimations, while the US PROMIS CAT software uses Expected A Priori (EAP). However, EAP pulls theta estimates towards the center of the population distribution, which may introduce bias in people with extreme responses [55]. For example, we found that respondents with the highest possible scores on almost all items were given a *T*-score of about 68 with EAP, where a *T*-score of 80 or 90 would have been more appropriate. Similar findings were found for the Dutch-Flemish PROMIS V1.0 Anxiety item bank [60]. With ML, however, a theta for participants with response patterns that exclusively comprise extreme responses cannot be estimated at all, which also creates a problem. In this study, we set the possible scale boundaries in the CAT simulation to −4 to 4, whereby people who score 1 or 5 on all CAT items get a theta score of −4 or 4, which is equivalent to a *T*-score of 10 and 90, respectively. Another possibility, used for the Dutch-Flemish PROMIS Anxiety item bank [60], is to use another estimation method, such as maximum a posteriori to estimate thetas for participants with extreme responses [47]. More research is recommended to find the optimal approach to estimate theta scores.

An interesting observation in this study was that the average *T*-score of the Satisfaction item bank was 47.5, while an average of 50.0 was to be expected in a general population sample (the average *T*-score of the Ability item bank was 50.6). This may indicate that Dutch people are, on average, less satisfied with their level of participation than US people, but further analyses are needed to explore possible alternative explanations.

In conclusion, the Dutch-Flemish PROMIS item banks Ability to Participate in Social Roles and Activities and Satisfaction with Social Roles and Activities showed sufficient psychometric properties in the general Dutch population. However, test–retest reliability and responsiveness need to be assessed in future studies. These item banks are now ready for use as CAT in research and clinical practice and will be made available through the Dutch-Flemish Assessment Center (http://www.dutchflemishpromis.nl). PROMIS CATs allow reliable and valid measurement of participation in an efficient and user-friendly way with limited administration time.

Acknowledgements The data collection for this project was financially supported by the Department of Epidemiology and Biostatistics of the VU University Medical Center, Amsterdam, the Netherlands. The Dutch-Flemish PROMIS group is an initiative that aims to translate and implement PROMIS item banks and CATs in the Netherlands and Flanders (http://www.dutchflemishpromis.nl). The Dutch-Flemish translation of the PROMIS item banks was supported by a grant from the Dutch Arthritis Association. We would like thank Michiel Luijten and Oguzhan Ogreden for their help with the CAT simulations.

Compliance with ethical standards

Conflict of interest Dr. C.B. Terwee is president of the PROMIS Health Organization. All authors are members of the Dutch-Flemish PROMIS group. All authors have no financial or non-financial conflicts of interest.

Ethical approval As this study did not involve experiments with patients, it was exempt from ethical approval according to the Dutch Medical Research in Human Subjects Act (WMO).

Human and animal rights All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and national research committees and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent According to the Dutch Medical Research in Human Subjects Act (WMO), obtaining informed consent was not necessary.

References

1. Deeg, D. J., & Bath, P. A. (2003). Self-rated health, gender, and mortality in older persons: Introduction to a special section. *Gerontologist, 43*(3), 369–371.
2. Douglas, H., Georgiou, A., & Westbrook, J. (2016). Social participation as an indicator of successful aging: An overview of concepts and their associations with health. *Australian Health Review, 41*(4), 455–462.
3. World Health Organization. (2001). *International classification of functioning, disability and health*. Geneva: Worl Health Organization.
4. Desrosiers, J., Noreau, L., & Rochette, A. (2004). Social participation of older adults in Quebec. *Aging Clinical and Experimantal Research, 16*(5), 406–412.
5. Obembe, A. O., & Eng, J. J. (2016). Rehabilitation interventions for improving social participation after stroke: A systematic review and meta-analysis. *Neurorehabilitation and Neural Repair, 30*(4), 384–392.
6. Powell, J. M., Rich, T. J., & Wise, E. K. (2016). Effectiveness of occupation- and activity-based interventions to improve everyday activities and social participation for people with traumatic brain injury: A systematic review. *American Journal of Occupational Therapy, 70*(3), 7003180040p1–7003180049p9.
7. Tanner, K., Hand, B. N., O'Toole, G., & Lane, A. E. (2015). Effectiveness of interventions to improve social participation, play, leisure, and restricted and repetitive behaviors in people with autism spectrum disorder: A systematic review. *American Journal of Occupational Therapy, 69*(5), 6905180010p1–6905180012p12.
8. Seekins, T., Shunkamolah, W., Bertsche, M., Cowart, C., Summers, J. A., Reichard, A., & White, G. (2012). A systematic scoping review of measures of participation in disability and rehabilitation research: A preliminary report of findings. *Disability and Health Journal, 5*(4), 224–232.

9. Taylor, A. M., Phillips, K., Patel, K. V., Turk, D. C., Dworkin, R. H., Beaton, D., et al. (2016). Assessment of physical function and participation in chronic pain clinical trials: IMMPACT/OMERACT recommendations. *Pain, 157*(9), 1836–1850.
10. Eyssen, I. C., Steultjens, M. P., Dekker, J., & Terwee, C. B. (2011). A systematic review of instruments assessing participation: Challenges in defining participation. *Archives of Physical Medicine and Rehabilitation, 92*(6), 983–997.
11. Magasi, S., & Post, M. W. (2010). A comparative review of contemporary participation measures' psychometric properties and content coverage. *Archives of Physical Medicine and Rehabilitation, 91*(9 Suppl), S17–S28.
12. Noonan, V. K., Kopec, J. A., Noreau, L., Singer, J., & Dvorak, M. F. (2009). A review of participation instruments based on the International Classification of Functioning, Disability and Health. *Disability and Rehabilitation, 31*(23), 1883–1901.
13. Phillips, N. M., Street, M., & Haesler, E. (2016). A systematic review of reliable and valid tools for the measurement of patient participation in healthcare. *BMJ Quality and Safety, 25*(2), 110–117.
14. Rettke, H., Geschwindner, H. M., & van den Heuvel, W. J. (2015). Assessment of patient participation in physical rehabilitation activities: An integrative review. *Rehabilitation Nursing, 40*(4), 209–223.
15. Stevelink, S. A., & van Brakel, W. H. (2013). The cross-cultural equivalence of participation instruments: A systematic review. *Disability and Rehabilitation, 35*(15), 1256–1268.
16. Tse, T., Douglas, J., Lentin, P., & Carey, L. (2013). Measuring participation after stroke: A review of frequently used tools. *Archives of Physical Medicine and Rehabilitation, 94*(1), 177–192.
17. Vergauwen, K., Huijnen, I. P., Kos, D., Van de Velde, D., van Eupen, I., & Meeus, M. (2015). Assessment of activity limitations and participation restrictions with persons with chronic fatigue syndrome: A systematic review. *Disability and Rehabilitation, 37*(19), 1706–1716.
18. Hahn, E. A., Devellis, R. F., Bode, R. K., Garcia, S. F., Castel, L. D., Eisen, S. V., et al. (2010). Measuring social health in the patient-reported outcomes measurement information system (PROMIS): Item bank development and testing. *Quality of Life Research, 19*(7), 1035–1044.
19. Hahn, E. A., DeWalt, D. A., Bode, R. K., Garcia, S. F., DeVellis, R. F., Correia, H., & Cella, D. (2014). New English and Spanish social health measures will facilitate evaluating health determinants. *Health Psychology, 33*(5), 490–499.
20. Cook, K. F., O'Malley, K. J., & Roddey, T. S. (2005). Dynamic assessment of health outcomes: Time to let the CAT out of the bag? *Health Services Research, 40*(5 Pt 2), 1694–1711.
21. 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*(8), 1759–1764.
22. PROMIS adult profile instruments. http://www.healthmeasures.net/images/promis/manuals/PROMIS_Profile_Scoring_Manual.pdf.
23. Hartman, J. D., & Craig, B. M. (2017). Comparing and transforming PROMIS utility values to the EQ-5D. *Quality of Life Research, 27*(3), 725–733.
24. Hanmer, J., Cella, D., Feeny, D., Fischhoff, B., Hays, R. D., Hess, R., et al. (2017). Selection of key health domains from PROMIS((R)) for a generic preference-based scoring system. *Quality of Life Research, 26*(12), 3377–3385.
25. Salisbury, C. (2012). Multimorbidity: Redesigning health care for people who use it. *The Lancet, 380*(9836), 7–9.
26. van Oostrom, S. H., Gijsen, R., Stirbu, I., Korevaar, J. C., Schellevis, F. G., Picavet, H. S., & Hoeymans, N. (2016). Time trends in prevalence of chronic diseases and multimorbidity not only due to aging: Data from general practices and health surveys. *PLoS ONE, 11*(8), e0160264.
27. Violan, C., Foguet-Boreu, Q., Flores-Mateo, G., Salisbury, C., Blom, J., Freitag, M., Glynn, L., Muth, C., & Valderas, J. M. (2014). Prevalence, determinants and patterns of multimorbidity in primary care: A systematic review of observational studies. *PLoS ONE, 9*(7), e102149.
28. Witter, J. P. (2016). The promise of patient-reported outcomes measurement information system-turning theory into reality: A uniform approach to patient-reported outcomes across rheumatic diseases. *Rheumatic Disease Clinics of North America, 42*(2), 377–394.
29. Purvis, T. E., Andreou, E., Neuman, B. J., Riley, L. H. III, & Skolasky, R. L. (2017). Concurrent validity and responsiveness of PROMIS health domains among patients presenting for anterior cervical spine surgery. *Spine (Phila Pa 1976), 42*(23), E1357–E1365.
30. Hinchcliff, M., Beaumont, J. L., Thavarajah, K., Varga, J., Chung, A., Podlusky, S., et al. (2011). Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item Health Profile and Functional Assessment of Chronic Illness Therapy-Dyspnea short form. *Arthritis Care & Research (Hoboken), 63*(11), 1620–1628.
31. Hinchcliff, M. E., Beaumont, J. L., Carns, M. A., Podlusky, S., Thavarajah, K., Varga, J., et al. (2015). Longitudinal evaluation of PROMIS-29 and FACIT-dyspnea short forms in systemic sclerosis. *Journal of Rheumatology, 42*(1), 64–72.
32. Katz, P., Pedro, S., & Michaud, K. (2017). Performance of the patient-reported outcomes measurement information system 29-item profile in rheumatoid arthritis, osteoarthritis, fibromyalgia, and systemic lupus erythematosus. *Arthritis Care Research (Hoboken), 69*(9), 1312–1321.
33. Lai, J. S., Beaumont, J. L., Jensen, S. E., Kaiser, K., Van Brunt, D. L., Kao, A. H., & Chen, S. Y. (2017). An evaluation of health-related quality of life in patients with systemic lupus erythematosus using PROMIS and Neuro-QoL. *Clinical Rheumatology, 36*(3), 555–562.
34. Yount, S. E., Beaumont, J. L., Chen, S. Y., Kaiser, K., Wortman, K., Van Brunt, D. L., et al. (2016). Health-related quality of life in patients with idiopathic pulmonary fibrosis. *Lung, 194*(2), 227–234.
35. Terwee, C. B., Roorda, L. D., de Vet, H. C., Dekker, J., van Westhovens, R., Cella, L. J., et al. (2014). Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Quality of Life Research, 23*(6), 1733–1741.
36. Statline, Bevoling; kerncijfers. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?ts=1536217517335. Accessed 29 Aug 2017.
37. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (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*(5 Suppl 1), S22–S31.
38. Rosseel, Y. (2012). An R package for structual equation modeling. *Journal of Statistical Software, 48*, 1–36.
39. Muthén, B. O., de Toit, S. H. C., Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. https://www.statmodel.com/download/Article_075.pdf. Accessed 7 Nov 2017.

40. Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.
41. Revelle, W. (2017). *psych: Procedures for personality and psychological research*. Evanston: Northwestern University.
42. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*, 1–55.
43. Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246.
44. Van der Ark, L. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1–19.
45. Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. Berlin: De Gruyer
46. Chalmers, P. (2012). A multidimensional Item Response Theory package for the R environment. *Journal of Statistical Software, 48*, 1–29.
47. Embretsen, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. New York: Psychology Press.
48. Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous Item Response Theory models. *Applied Psychological Measurement, 27*, 289–298.
49. McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49–57.
50. Hortensius, L. (2012). *Advanced measurement: Logistic regression for DIF analysis*. Minneapolis, MN: University of Minnesota.
51. Devellis, R. (2016). PROMIS 1 social supplement. Retrieved from https://dataverse.harvard.edu/dataverse.xhtml?alias=HealthMeasures.
52. 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.
53. Cappelleri, J. C., Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and Item Response Theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648–662.
54. Magis, D. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*, 1–31.
55. Smits, N. (2016). On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: A simulation study. *Quality of Life Research, 25*(7), 1635–1644.
56. Crins, M. H., Roorda, L. D., Smits, N., de Vet, H. C., Westhovens, R., Cella, D., et al. (2015). Calibration and Validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. *PLoS ONE, 10*(7), e0134094.
57. Crins, M. H., Roorda, L. D., Smits, N., de Vet, H. C., Westhovens, R., Cella, D., et al. (2016). Calibration of the Dutch-Flemish PROMIS pain behavior item bank in patients with chronic pain. *European Journal of Pain, 20*(2), 284–296.
58. Crins, M. H. P., Terwee, C. B., Klausch, T., Smits, N., de Vet, H. C. W., Westhovens, R., et al. (2017). The Dutch-Flemish PROMIS physical function item bank exhibited strong psychometric properties in patients with chronic pain. *Journal of Clinical Epidemiology, 87*, 47–58.
59. 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 and the Health Professions, 40*(1), 79–105.
60. Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2017). Development of a computerized adaptive test for anxiety based on the Dutch-Flemish version of the PROMIS Item Bank. *Assessment*, 1073191117746742.
61. Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*(4), 447–460.
62. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educational and Psychological Measurement, 73*, 5–26.
63. Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales. A practical guide to their development and use*. New York: Oxford University Press.