Towards standardization of fatigue measurement: Psychometric properties and reference values of the PROMIS Fatigue item bank in the Dutch general population

Caroline B Terwee1*, Ellen BM Elsman1 and Leo D Roorda2

Abstract

Background: There is little consensus on how to measure fatigue.

Objectives: To standardize the measurement of fatigue across populations, we aimed to assess the psychometric properties of the PROMIS Fatigue item bank in the Dutch general population and obtain reference values.

Methods: A sample of 1006 people participating in an internet panel completed the full v1.0 PROMIS Fatigue item bank (95 items). Structural validity (item response theory (IRT) assumptions and IRT model fit), measurement invariance/cross-cultural validity (absence of differential items functioning (DIF) for demographic variables and language, compared to data from US participants in PROMIS wave 1), and (internal) reliability (percentage of respondents with reliable estimates) were assessed.

Results: The IRT model assumptions were considered met (ECV 0.86, Omega-H 0.92), all items fitted the IRT model, no items showed DIF for demographic variables and seven for language, but with negligible impact on T-scores. Reliable fatigue T-scores were found for 98.3%, 69.8–82.6%, and 96.5% of the respondents with the full item bank, the standard short forms, and a simulated computerized adaptive test (CAT), respectively. The CAT administered on average only five items. A T-score of 49.1 represented the average score of the Dutch general population, T-scores <55 are considered within normal limits, T-scores of 55–59 indicate mild fatigue, T-scores of 60–70 indicate moderate fatigue, and T-scores >70 indicate severe fatigue.

Conclusions: The PROMIS Fatigue item bank showed sufficient structural validity, no measurement invariance for demographic characteristics, sufficient cross-cultural validity, and sufficient (internal) reliability in the Dutch general population.

Keywords
patient-reported outcomes, cross-cultural validity, item response theory, PROMIS, reference values

Introduction

Fatigue is a common symptom in multiple conditions, such as cancer,1 cardiovascular disease,2,3 chronic obstructive pulmonary disease (COPD),4 inflammatory bowel disease,5 skin disease,6 multiple sclerosis,7 rheumatoid arthritis,8 and many others. It has been included as one of the core outcomes, that is, outcomes that matter most to patients, in about one third of the Standard Sets developed by the International Consortium for Health Outcomes Measurement (ICHOM).9

Despite the importance of fatigue, there is little consensus on how to measure it. Numerous generic and disease-specific
patient-reported outcome measures (PROMs) exist to measure fatigue. For example, systematic reviews identified 25 PROMs for measuring cancer-related fatigue,10 43 PROMs for measuring fatigue in hemodialysis patients,11 10 PROMs for measuring fatigue in non-cancer gastrointestinal disorders,12 and 31 fatigue questionnaires for multiple sclerosis, Parkinson’s disease and stroke.13 In nine ICHOM Standard Sets recommending the measurement of fatigue, six different PROMs were suggested.9 The available fatigue questionnaires differ in content and quality (i.e., psychometric properties) and total scores are not comparable, hindering benchmarking and quality of care improvements.

The severity and impact of fatigue on daily activities, should be measured with instruments that have sufficient psychometric properties, including validity (content, structural, construct, and cross-cultural validity), reliability (internal consistency, test–retest reliability, and measurement error), responsiveness, interpretability, and low completion burden for patients. Furthermore, the measurement of fatigue should, wherever possible, be standardized in research and clinical practice, in order to enable comparison of the burden of disease and treatment within and across populations.

To improve the quality of fatigue measurement and standardize its measurement across populations, the Patient-Reported Outcomes Measurement Information System (PROMIS)® initiative developed a highly precise and universal applicable (or generic) fatigue PROM that can be used in healthy persons as well as patients with varying medical conditions. The PROMIS Fatigue measure was built on items from existing PROMs that had undergone testing previously, identified in an extensive literature search as well as focus groups with a mixed sample of patients.14 In addition, cognitive interviews were performed with patients with a diverse range of chronic health conditions.15 Using a modern psychometric technique (item response theory (IRT)) an “item bank” of 95 fatigue items was created, measuring a range of self-reported symptoms, from mild subjective feelings of tiredness to an overwhelming, debilitating, and sustained sense of exhaustion that likely decreases one’s ability to execute daily activities and function normally in family or social roles.16 With IRT analyses items in an item bank are ordered on a scale, according to the fatigue level they address (also called item “location” or item “difficulty”). For example, the item “How often did you have enough energy to exercise strenuously?” indicates a low level of fatigue because even patients with a little fatigue may answer “sometimes,” while the item “How often were you too tired to watch television?” indicates a high level of fatigue because only patients with high levels of fatigue will answer “sometimes.” Each item has its unique location on the scale and also a unique discriminative ability.17 Once the item locations and discriminative abilities are defined, fixed subsets of items can be administered to patients as short forms (standard short forms of 4, 6, 7, and 8 items were developed), or the item bank can be administered as computerized adaptive test (CAT). In a CAT items are selected from the item bank by a computer based on a person’s responses to previous items.18 Scores of short forms and CAT are computed taking the item location and discriminative ability of the items into account. Scores of short forms and CAT are on the same scale (or metric), which makes them comparable.

Research supports the psychometric properties of the generic PROMIS Fatigue measures in the general population and across varying conditions. One psychometric property, content validity, of the PROMIS Fatigue item bank was supported in patients with rheumatoid arthritis and multiple sclerosis19,20 Other psychometric properties, internal consistency, structural validity, test–retest reliability, construct validity, and responsiveness, of different PROMIS Fatigue short forms and CAT were supported across patient populations with a wide range of conditions, such as rheumatologic conditions, back pain, Myalgic Encephalomyelitis/Chronic Fatigue Syndrome, cancer, HIV, chronic heart failure, COPD, depression, and others.21–36 Egerton et al. evaluated the measurement properties of self-report questionnaires for measuring fatigue in older people. PROMIS Fatigue item bank and short forms performed best out of 77 identified questionnaires.37 Because of the innovative psychometric methods used to develop item banks and its universal applicability, PROMIS CATs may be able to replace disease-specific PROMs. Since a CAT selects items that are most informative for each patient, reliability and responsiveness are high. PROMIS CATs have found to be as responsive as disease-specific PROMs.34,36,38,39

However, all of these studies so far, addressing the development and psychometric properties of the PROMIS fatigue item bank in multiple populations, were performed in the US. There is no evidence yet for the psychometric properties of the PROMIS Fatigue measures outside of the US. There is also limited evidence for measurement invariance across demographic variables and across countries (cross-cultural validity), which is important because item parameters may be different across countries, which could impact scores and hinder comparisons between groups differing with respect to demographic variables or cultural background. The aim of this study was therefore to assess the psychometric properties structural validity, measurement invariance/cross-cultural validity, and (internal) reliability of the Dutch-Flemish version of the v1.0 PROMIS Fatigue item bank in the Dutch general population, and to assess Dutch reference values, to facilitate large-scale international implementation of this item bank as short form or CAT in research and clinical practice.

Methods
The Medical Ethical Committee of Amsterdam UMC, location VUmc, the Netherlands, confirmed that the study
protocol was exempted from ethical approval according to the Dutch Medical Research in Human Subjects Act (WMO), as no experiments were conducted. The study adhered to the tenets of the Declaration of Helsinki.

Participants and procedures

A cross-sectional study was performed. A data collection company (Desan Research Solutions) recruited people of the Dutch general population from an existing internet panel in 2016 (more details about the panel can be found here). We considered a sample of at least 1000 people sufficient for item parameter estimation. The study sample was selected to be representative for the Dutch general population with respect to age distribution (18–40; 40–65; >65), gender, educational level (low, middle, high), region of residence (north, east, south, west) and ethnicity (native Dutch, first- and second-generation western immigrant, first- and second-generation non-western immigrant). We compared the characteristics of the participants to data from Statistics Netherlands in 2016 to check for a maximum allowable deviation of 2.5% per variable.

Measures

A web-based survey was used, in which skipping items was not allowed. Participants completed the full v1.0 PROMIS Fatigue item bank, consisting of 95 items, or more specifically statements or questions referring to the severity or impact of fatigue. All items have five response options, higher scores indicate more fatigue, except for eight items referring to having energy to do things, which were recoded. Example items and response options are provided in Box 1. The recall period is the past 7 days. Additionally, participants completed questions regarding sociodemographic characteristics (age, gender, education, region of residence, and ethnicity).

Statistical analyses

Details of all statistical methods and their criteria are presented in Table 1. A summary of the analyses is presented below.

Structural validity. First we checked data assumptions required for IRT modeling. We checked whether the item bank was unidimensional enough for IRT analysis (i.e., measuring only one construct), by using confirmatory and bifactor analyses. We also evaluated local independence by checking whether residual correlations among the items were not too high. We finally checked the monotonicity assumption, which states that the probability for patients to select higher response categories should increase with increasing levels of fatigue. After assuring that the assumptions were met, we fitted an IRT model (Graded Response Model) to the response data, estimated the IRT item parameters (i.e., item locations/thresholds and item discrimination parameters), and assessed the fit of each item to the model.

Measurement invariance/cross-cultural validity. We examined whether people from different subgroups (e.g., males versus females) with the same level of the fatigue have similar probabilities of giving a certain response to an item (measurement invariance). If that is the case, the same IRT parameters can be used to calculate and compare scores across groups. We evaluated measurement invariance for age, gender, education, region, and ethnicity, by comparing a series of ordinal regression models, assessing whether, when controlling for the level of fatigue, the probability of giving a certain response to an item is the same across groups. We also evaluated measurement invariance for language (Dutch vs American-English), which can be considered evidence for cross-cultural validity. For the latter aim, we compared our sample to a sample of 21,133 individuals from the US general population that was used for developing the original item bank (PROMIS Wave 1, obtained from the HealthMeasures Dataverse repository). PROMIS Wave 1 data were collected in 2006–2007 by a polling firm. Data consisted of 7005 individuals who completed the full PROMIS Fatigue item bank and 14,128 individuals who completed 7 items measuring fatigue experience, 7 measuring fatigue impact, and also 7 items from each of the other 12 domains included in PROMIS wave 1 testing. Mean (SD) age of the sample was 53.1 (17.1) and 52% were women.

Reliability. To evaluate reliability, first fatigue scores were calculated for all study participants based on the PROMIS Fatigue full item bank, derivative short forms (4a, 6a, 8a, and 7a) and a simulated CAT. PROMIS scores are, by
| Property                  | Analysis                  | Description of analysis                                                                                                                                                                                                                                                                                                                                 | Indices                                                                 | Criteria for good properties | Software   |
|--------------------------|---------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|------------------------------|------------|
| IRT model assumptions    |                           |                                                                                                                                                                                                                                                                                                                                                         | Scaled Comparative fit Index (CFI)                                      | >0.95                        | R-package  |
| Unidimensionality        | Confirmatory factor analysis (CFA) | In a one-factor model all items are considered to load on only one factor (underlying construct/trait). A one-factor model was fitted on the polychoric correlation matrix with Weighted Least Squares with Mean and Variance adjustment (WLSMV) estimation | Scaled Tucker Lewis Index (TLI)                                         | >0.95                        | Lavaan     |
|                          |                           |                                                                                                                                                                                                                                                                                                                                                         | Scaled Root Mean Square error of Approximation (RMSEA)                  | <0.06                        | Version 0.6–5 |
|                          |                           |                                                                                                                                                                                                                                                                                                                                                         | Standardized Root Mean Residual (SRMR)                                 | <0.08                        |            |
| Bi-factor analysis       |                           | In a bi-factor model, all items are considered to load on one general factor. In addition, items can load on group factors that capture item covariation that is independent of the covariation due to the general factor. An exploratory bi-factor analysis was performed with one general factor and three group factors. Indices indicate the relative strength of the general factor in relation to the group factors | Explained Common Variance (ECV)                                        | >0.60                        |            |
|                          |                           |                                                                                                                                                                                                                                                                                                                                                         | Omega-H                                                                  | >0.80                        |            |
| Local independence       | CFA                       | After controlling for the dominant factor, there should be no important covariance among item responses. Residual correlations between the items in the one-factor CFA were examined. The impact of local dependence was tested by estimating the maximum change in item parameters if items with local dependence would be removed from the item bank | Correlation coefficient                                                  | <0.20                        |            |
| Monotonicity             | Mokken scaling            | The probability of endorsing a higher item response category should increase (or at least not decrease) with increasing levels of the underlying construct/trait. Monotonicity was evaluated by fitting a non-parametric IRT model, with Mokken scaling | Scale - Scalability coefficient H Items - Scalability coefficient H      | >0.50                        | R-package  |
|                          |                           |                                                                                                                                                                                                                                                                                                                                                         |                                                                          | >0.30                        | Mokken     |
|                          |                           |                                                                                                                                                                                                                                                                                                                                                         |                                                                          |                              | Version 2.8.11 |

Table 1. Psychometric properties studied, criteria, software, or R-packages used.

(continued)
Table 1. (continued)

| Property            | Analysis             | Description of analysis                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|---------------------|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Item fit            | IRT modeling        | A logistic Graded Response Model (GRM) using the Bock–Aitkin maximal likelihood estimation was used. Two parameters were estimated for each item: Item thresholds locate the items along the measured trait. Item slopes refers to the discriminative ability of the items. To assess the fit of each item to the GRM model we used a generalization of Orlando and Thissen’s S-X² for polytomous data. |
| Measurement invariance | No differential item functioning (DIF) | Ordinal logistic regression analyses People from different groups with the same level of the construct/trait should have the same probability of giving a certain response to an item. If these probabilities are not the same, there is DIF. Uniform DIF exists when the magnitude of the DIF is consistent across the entire range of the trait (i.e., differences in item thresholds). Non-uniform DIF exists when the magnitude or direction of DIF differs across the trait (i.e., differences in item slopes). We evaluated DIF for age (18–40; 40–65; >65), gender (male, female), educational level (low, middle, high), region of residence (north, east, south, west) and ethnicity (native Dutch, first- and second-generation western immigrant, first- and second-generation non-western immigrant). Three ordinal logistic regression models were compared, regressing the item response on the trait level (model 1), trait level plus group factor (model 2), and trait level, group factor and interaction of trait level and group factor (model 3). |

Change in Mcfadden $R^2$ between models 1 and 2 (uniform DIF) and between models 2 and 3 (non-uniform DIF) $<0.02$ R-package Lordif Version 0.3–3

S-X² and p-value $>0.001$ R-package Mirt Version 1.31

Cross-cultural validity (continued)
default, based on the item parameters of the original IRT model of the US calibration sample on which the item bank was developed (unless large measurement invariance is found), so that scores are comparable across populations and countries. IRT-based scores always have an average of 0 and SD of 1 in the calibration sample (theta scale). PROMIS, however, uses a T-score metric, which is obtained by multiplying the theta score by 10 and adding 50. T-scores of almost all PROMIS domains thereby have a mean of 50 and a standard deviation of 10 in the US reference population. PROMIS T-scores can be calculated from the raw item scores using the online HealthMeasures Scoring Service program, provided by the US Assessment Center.

However, for the CAT simulation, we needed the original

---

**Table 1.** (continued)

| Property          | Analysis                          | Description of analysis                                                                                                                                                                                                 | Indices       | Criteria for good properties | Software            |
|-------------------|-----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|-----------------------------|---------------------|
| No DIF            | Ordinal logistic regression analyses | See above. We evaluated DIF for language (English, Dutch)                                                                                                                                                               | See above     | <0.02                       | R-package Lordif Version 0.3–3 |
| Reliability       | Percentage reliable scores        | SE (T-score)                                                                                                                                                                                                              | SE (T-score)  | <3.16                       | R-package catR Version 3.16 |

*Table 2. Sociodemographic characteristics of the study participants and the Dutch general population.*

| Sociodemographic characteristic | Study participants* (n = 1006) | Dutch adult population 2016† (n = 13.6 million) |
|---------------------------------|---------------------------------|-----------------------------------------------|
| Age in years, mean ± SD (range)  | 52 ± 17 (18–93)                 | 33.7                                          |
| 18–39                           | 32.5                            | 32.7                                          |
| 40–64                           | 43.7                            | 43.6                                          |
| ≥65                             | 23.8                            | 22.7                                          |
| Gender                          |                                 |                                               |
| Male                            | 46.9                            | 49.2                                          |
| Female                          | 53.1                            | 50.8                                          |
| Educational level               |                                 |                                               |
| Low                             | 27.8                            | 30.2                                          |
| Middle                          | 41.2                            | 40.2                                          |
| High                            | 31.0                            | 29.6                                          |
| Region of residence             |                                 |                                               |
| North                           | 11.5                            | 10.2                                          |
| East                            | 19.6                            | 20.8                                          |
| South                           | 23.4                            | 21.6                                          |
| West                            | 45.5                            | 47.4                                          |
| Ethnicity                       |                                 |                                               |
| Native                          | 79.3                            | 78.6                                          |
| 1st and 2nd generation western immigrant | 10.9                             | 10.3                                          |
| 1st and 2nd generation non-western immigrant | 9.7                             | 11.2                                          |

*All results expressed as % unless otherwise noted.

SD: standard deviation;

†Based on data from statistics Netherlands (https://www.cbs.nl).
US item parameters, which were obtained from the HealthMeasures group.\textsuperscript{46}

Reliability (or precision) within IRT is inversely related to the standard error (SE) of the estimated fatigue score (this form of reliability is also called internal reliability or internal consistency because it is based on only one measurement). Each score is associated with a SE. The SE differs across the scale, and is usually lower in the middle of the scale than at the ends of the scale.\textsuperscript{17,47} We calculated the number of participants who got a reliable score on the T-score scale (SE<3.16, which equals a reliability of 0.90) with the full bank, short forms, and a simulated CAT. We simulated a CAT using the standard PROMIS CAT start and stopping rules. The start item was the item that is most informative (i.e., best reliability) for people with an average level of fatigue, which is item FATIMP3 (“How often did you have to push yourself to get things done because of your fatigue?”). A minimum of 4 items were administered and the CAT stopped when a SE of three on the T-score metric was reached or a maximum of 12 items were administered. We also plotted the SE across T-scores for the entire item bank, the standard short forms and the simulated CAT.

Table 3. IRT assumptions and model fit of the V1.0 PROMIS Fatigue Item Bank.

| Property                        | Indices                                      | Criterion | Result          |
|---------------------------------|----------------------------------------------|-----------|-----------------|
| IRT model assumptions           |                                              |           |                 |
| Unidimensionality               | Scaled Comparative fit Index (CFI)           | >0.95     | 0.955           |
|                                 | Scaled Tucker–Lewis Index (TLI)              | >0.95     | 0.954           |
|                                 | Scaled Root Mean Square Error of Approximation (RMSEA) | <0.06 | 0.075           |
|                                 | Standardized Root Mean Residual (SRMR)       | <0.08     | 0.046           |
|                                 | Explained Common Variance (ECV)              | >0.60     | 0.86            |
|                                 | Omega-H                                      | >0.80     | 0.92            |
| Local independence              | Percentage item pairs with local independence | <0.20     | 99.7%           |
| Monotonicity                    | Scale - Scalability coefficient $H$          | >0.50     | 0.71            |
|                                 | Items - Scalability coefficient $H_i$        | >0.30     | 0.43–0.77       |
| IRT model fit                   |                                              |           |                 |
| Item fit                        | Number of items with S-X$^2$ p-value > 0.001 | >0.001    | 95              |
| Item parameters                 | Item locations/thresholds                    | —         | –2.54–3.35      |
|                                 | Item discrimination parameters               |           | 1.20–4.14       |

*All items showed uniform DIF.
\textsuperscript{b} = also DIF in Spanish language (results obtained from HealthMeasures, personal communication).

Table 4. Cross-cultural validity: items that showed differential item functioning in the Dutch versus American-English sample.

| Item     | Description                                                                 | $R^2_{\text{change}}$ |
|----------|------------------------------------------------------------------------------|------------------------|
| FATEXP42 | How much mental energy did you have on average?                              | 0.0305                 |
| ANS      | I have energy                                                                 | 0.0254                 |
| FATIMP28b| How hard was it for you to carry on a conversation because of your fatigue?  | 0.0245                 |
| AN2b     | I feel tired                                                                  | 0.0241                 |
| FATIMP25 | How often was it an effort to carry on a conversation because of your fatigue?| 0.0240                 |
| FATEXP2b | How often did you feel run-down?                                             | 0.0221                 |
| H17b     | I feel fatigued                                                               | 0.0207                 |

Table 5. Reliability of the V1.0 PROMIS Fatigue full bank, short forms, and CAT.

| Measure   | Percentage of respondents reliably ($r > 0.90$) estimated |
|-----------|----------------------------------------------------------|
| Full bank | 98.3                                                     |
| CAT       | 96.5                                                     |
| Short form 8a | 82.6                                                    |
| Short form 6a | 79.8                                                    |
| Short form 4a | 75.8                                                    |
| Short form 7a | 69.8                                                    |
To obtain Dutch reference values, we calculated the mean (SD) T-score for the entire group of study participants, and for age-range (18–34 years, 35–44 years, 45–54 years, 55–64 years, 65–74 years, and ≥75 years) and gender subpopulations. We also calculated fatigue scores of 0.5*SD, 1*SD, and 2*SD above the average of the general population as thresholds for mild, moderate and severe fatigue respectively.

**Results**

A sample of 1006 individuals completed the online study questionnaire (mean age 52 (SD 17), 53% female) between July and November 2016. All participants had complete data. The demographic characteristics of the participants are summarized and compared to the Dutch population in 2016 in Table 2. All differences were less than the 2.5% agreed upon.

**Structural validity**

The IRT model assumptions were considered met. With respect to unidimensionality, the RMSEA was slightly too high (0.075 instead of <0.06) but the high ECV (0.86) and Omega-H (0.92) indicated that the item bank was unidimensional enough for performing IRT analysis. The assumption of local independence was also considered met as only 28/8930 (0.03%) of item pairs showed a residual correlation >0.20 (range 0.20–0.47). Moreover, the maximum possible impact of local dependence on the item parameter estimations was very small (maximum impact on discrimination parameter 0.05, maximum impact on thresholds 0.04). The assumption of monotonicity was also met (Table 3). All items fitted the IRT model. The item thresholds ranged from −2.54 to 3.35, which means that the item bank can measure a broad range of fatigue, from people with about 2.5*SD less fatigue than average to about 3*SD more fatigue than average. Discrimination parameters ranged from 1.2 to 4.1.

**Measurement invariance/cross-cultural validity**

No items showed DIF for age, gender, education, region or ethnicity. Seven items showed uniform DIF for language. Dutch persons with similar levels of fatigue as US persons, were more inclined to respond that they are tired on these items (Table 4). However, the impact of DIF on the total score was negligible (Supplement 1).

**Reliability**

In total, 98.3% of the respondents had reliable (r>0.90) fatigue scores with the full item bank, 69.8–82.6% with the short forms, and 96.5% with the CAT (Table 5). With CAT the mean number of items administered was 5 and 777 individuals (77.2%) completed a maximum of five items. Thirty-six individuals (3.6%) completed 12 items.

**Dutch reference values**

The average fatigue score of the Dutch general population was 49.1 (SD 10.8) (Table 6). Males had slightly lower fatigue levels than females (47.5 (10.7) versus 50.4 (10.8)). Respondents in the age group 35–44 had the highest fatigue level (50.6 (10.3)), while the average level of fatigue decreased by age after the age of 44 to an average of 45.4 (10.3) at the age of 75+. The following interpretation thresholds were defined: <55 = within normal limits, 55–59 = mild, 60–70 = moderate, and >70 = severe fatigue.

| Table 6. PROMIS fatigue reference valuesa for the Dutch general population by age and gender, and comparisons with the US reference population. |
|---|---|---|---|
| | N Dutch population (%) | Dutch mean T-score (SD) | N US population (%) | US mean T-score (SD) |
| Total | 1006 (100) | 49.1 (10.8) | 3067 (100) | 50.0 (10.0) |
| Gender | | | | |
| Male | 472 (47) | 47.5 (10.7) | 1183 (39) | 48.2 (9.6) |
| Female | 534 (53) | 50.4 (10.8) | 1884 (61) | 51.1 (10.1) |
| Age in years | | | | |
| 18–34 | 192 (19) | 50.3 (9.6) | 706 (23) | 50.5 (9.7) |
| 35–44 | 229 (23) | 50.6 (10.3) | 551 (18) | 51.0 (10.7) |
| 45–54 | 120 (12) | 49.4 (11.2) | 513 (17) | 51.6 (10.1) |
| 55–64 | 192 (19) | 49.5 (11.7) | 516 (17) | 49.7 (10.8) |
| 65–74 | 173 (17) | 47.1 (11.2) | 396 (13) | 48.1 (9.3) |
| 75+ | 100 (10) | 45.4 (10.3) | 385 (13) | 48.0 (8.3) |

SD: standard deviation

aT-scores, higher scores represent more fatigue.
Discussion

The Dutch-Flemish v1.0 PROMIS Fatigue item bank showed sufficient structural validity, no measurement invariance for important demographic characteristics, sufficient cross-cultural validity, and (internal) reliability. With the full item bank 98.3% of the respondents had reliable ($r > 0.90$) fatigue scores. With the short forms this was 69.8–82.6% and with the CAT 96.5%, with on average only five items.

This is the first study that evaluated cross-cultural validity of the PROMIS Fatigue item bank, contributing to its international applicability. Seven items showed DIF for language, indicating that Dutch people with on average similar levels of fatigue as US people, are more inclined to respond to these items that they experience fatigue. For some items, this may be due to the translation. For example, item FATEXP2 “How often did you feel run-down?” was difficult to translate. In addition, we had difficulty to make two different translations for item AN2 (I feel tired) and H17 (I feel fatigued) because no distinction is made between tired and fatigued in the Dutch language. However, for other DIF items we did not find any problems with the translation. The magnitude of the DIF for all seven items was small. Of these seven items, only one item (H17) is included in the most often used standard 8a short form. The magnitude of the DIF of this item was low ($R^2 0.0207$, just slightly above the critical value of 0.02); therefore, the impact of DIF on the short form T-scores is expected to be negligible. Also, only one of the DIF items (AN5) was selected in the simulated CATs (in 13% of the participants). The magnitude of the DIF of this item was also quite low ($R^2 0.0254$), so the impact of DIF on the CAT T-scores is also expected to be very low.

We did not assess the presence of (chronic) conditions in our study sample, but considering a prevalence of chronic diseases of about 40% in the Dutch general population,49 we assume that our study sample included a large proportion of people with different conditions. Therefore, our study adds to the accumulating evidence that fatigue can be measured validly and reliably across patients with a wide range of conditions with generic PROMIS Fatigue measures.21–36 Previous research showed the relevance of the PROMIS Fatigue items across different patient populations.15,19,20 A study in rheumatoid arthritis patients also showed that most patients would not give a different response when asked about a general sense of fatigue compared to fatigue attributed to their disease.19

This body of evidence provides an important justification and encouragement for the standardization of patient-reported outcome measurement across medical conditions. It is too time-consuming and costly to build in many different PROMs in electronic health records, it is difficult for healthcare providers to use different PROMs in different setting or for patients with different conditions and interpret the results correctly, and it is burdensome for patients with multiple conditions to complete different PROMs for different healthcare providers.30–52 Standardization is needed for large-scale assessment, comparison of outcomes within and between patient groups, and improvement of the quality of the health care system and the health of patients. To facilitate the transition from using traditional PROMs to PROMIS, so called “crosswalks” can be created to transform scores of currently frequently used fatigue PROMs, such as the Modified Fatigue Impact Scale (MFIS), to the PROMIS Fatigue metric.53,54

Our study also showed that CAT is a very efficient and patient-friendly way of measuring outcomes. With CAT 96.5% of the patients got a reliable score with on average only five items. Moreover, CAT clearly outperformed the short forms. Other studies found similar results for other PROMIS item banks.55–62

This study additionally provided Dutch reference values for the PROMIS Fatigue measures. A T-score of 49.1 represents the average score of the Dutch general population, which is quite similar to the average T-score of 50 in the US population. Also the thresholds for mild, moderate, and severe fatigue of 55, 60, and 70, respectively, were found to be similar in the Dutch population as in the US population. Evidence on the minimal detectable change and minimal important change of PROMIS Fatigue measures is still scarce. Change scores of about 11–13 T-scores points have been found to be minimally detectable and change scores of about 2–4 T-score points have been found to be minimally important.63–66 However, these studies have methodological limitations and more high quality evidence is needed. Evidence on the psychometric properties of the PROMIS Fatigue measures in other countries is also required to enable comparison of outcomes of health care across countries.

A strength of this study is its large sample size and comparison to recent data from the Dutch general population, as well as comparison to a large US general population sample. A limitation is the lack of knowledge about the presence of (chronic) conditions in the study sample. Therefore, we were not able to evaluate measurement invariance for (chronic) conditions.

Conclusion

The Dutch-Flemish v1.0 PROMIS Fatigue item bank showed sufficient structural validity, no measurement invariance for important demographic characteristics, sufficient cross-cultural validity, and sufficient (internal) reliability in the Dutch general population. A T-score of 49.1 represents the average score of the Dutch general population, T-scores $<55$ are considered within normal limits, T-scores of $55–59$ indicate mild fatigue, T-scores of $60–70$ indicate moderate
fatigue, and T-scores >70 indicate severe fatigue. This study provides additional evidence for the universal applicability of the PROMIS fatigue item banks across populations differing with respect to demographic characteristics, it provides convincing evidence for its international applicability, it contributes to the interpretability of scores, and therewith provides evidence for the use of PROMIS as the international standard for measuring fatigue.

Acknowledgment
We would like to thank Michiel Luijten for help with some of the statistical analyses.

Availability of data and material
The dataset is available upon request from the corresponding author.

Ethics approval
The Medical Ethical Committee of Amsterdam UMC, location VUmc, the Netherlands, confirmed that the study protocol was exempted from ethical approval according to the Dutch Medical Research in Human Subjects Act (WMO), as no experiments were conducted.

Authors contributions
CB Terwee and LR Roorda designed the study and were responsible for the data collection. CB Terwee and EBM Elsman conducted the analyses. CB Terwee drafted the manuscript and all authors contributed to the writing and finally approved the manuscript.

Declaration of Conflicting Interests
CB Terwee is board member of the Dutch-Flemish PROMIS Organization. CB Terwee and LD Roorda are representatives of the Dutch-Flemish PROMIS National Center. EBM Elsman has nothing to declare.

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

ORCID iDs
Caroline B. Terwee https://orcid.org/0000-0003-4570-2826
Ellen Elsman https://orcid.org/0000-0002-0680-6430
Leo D. Roorda https://orcid.org/0000-0001-9348-5390

Supplemental Material
Supplemental material for this article is available online.

References
1. Al Maqbali M, Al Sinani M, Al Naamani Z, et al. Prevalence of fatigue in patients with cancer: a systematic review and meta-analysis. J Pain Symptom Manage 2021; 61: 167–189. DOI: 10.1016/j.jpainsymman.2020.07.037.
2. Casillas J-M, Damak S, Chauvet-Gelinier J-C, et al. Fatigue et maladies cardiovasculaires. Ann de Réadaptation de Méd Phys 2006; 49: 30939–319402. DOI: 10.1016/j.annrmp.2006.04.002.
3. Cuming TB, Packer M, Kramer SF, et al. The prevalence of fatigue after stroke: A systematic review and meta-analysis. Int J Stroke 2016; 11: 968–977. DOI: 10.1177/1747493016669861.
4. Goertz YMJ, Spruit MA, Van’t Hul AJ, et al. Fatigue is highly prevalent in patients with COPD and correlates poorly with the degree of airflow limitation. Ther Adv Respir Dis 2019; 13. DOI: 10.1177/1753466619878128.
5. Chavarría C, Casanova MJ, Chaparro M, et al. Prevalence and factors associated with fatigue in patients with inflammatory bowel disease: a multicentre study. J Crohn’s Colitis 2019; 13: 996–1002. DOI: 10.1093/jecoco/jjz024.
6. Misery L, Shourick J and Taieb C. Prevalence and characterization of fatigue in patients with skin diseases. Acta Dermato Venereol 2020; 100: adv03272020. DOI: 10.2340/00015555-3694.
7. Rooney S, Wood L, Moffat F, et al. Prevalence of fatigue and its association with clinical features in progressive and non-progressive forms of Multiple Sclerosis. Mult Sclerosis Related Disorders 2019; 28: 276–282. DOI: 10.1016/j.msard.2019.01.011.
8. Nikolou S, Bode C, Taal E, et al. Fatigue and factors related to fatigue in rheumatoid arthritis: a systematic review. Arthritis Care Res 2013; 65: 1128–1146. DOI: 10.1002acr.21949.
9. Terwee CB, Zuidergeest M, Vonkeman HE, et al. Common Patient-Reported Outcomes across ICHOM Standard Sets – the Potential Contribution of PROMIS, 2020.
10. Al Maqbali M, Hughes C, Gracey J, et al. Quality assessment criteria: psychometric properties of measurement tools for cancer related fatigue. Acta Oncol 2019; 58: 1286–1297. DOI: 10.1080/0284186x.2019.1622773.
11. Ju A, Unruh ML, Davison SN, et al. Patient-reported outcome measures for fatigue in patients on hemodialysis: a systematic review. Am J Kidney Dis 2018; 71: 327–343. DOI: 10.1053/j.ajkd.2017.08.019.
12. Jungyoun Han C, Heitkemper MM and Jarrett ME. Fatigue measures in noncancer gastrointestinal disorders. Gastroenterol Nurs 2016; 39: 443–456. DOI: 10.1097/sga.0000000000001747.
13. Elbers RG, Rieberg MB, van Wegen EEH, et al. Self-report fatigue questionnaires in multiple sclerosis, Parkinson’s disease and stroke: a systematic review of measurement properties. Qual Life Res 2012; 21: 925–944. DOI: 10.1007/s11136-011-0009-2.
14. DeWalt DA, Rothrock N, Yount S, et al. Evaluation of item candidates. Med Care 2007; 45: S12–S21. DOI: 10.1097/01.mlr.0000254567.79743.e2.

15. Christodoulou C, Junghaenel DU, DeWalt DA, et al. Cognitive interviewing in the evaluation of fatigue items: results from the patient-reported outcomes measurement information system (PROMIS). Qual Life Res 2008; 17: 1239–1246. DOI: 10.1007/s11136-008-9402-x.

16. Lai J-S, Cella D, Choi S, et al. How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example. Arch Phys Med Rehabil 2011; 92: S20–S27. DOI: 10.1016/j.apmr.2010.08.033.

17. Embretsen SE and Reise SP. Item response theory for psychologists. New York: Psychology Press, 2000.

18. Bjorner JB, Chang C-H, Thissen D, et al. Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res 2007; 16(Suppl 1): 95–108. DOI: 10.1007/s11136-007-9168-6.

19. Bartlett SJ, Gutierrez AK, Butanis A, et al. Combining online and in-person methods to evaluate the content validity of PROMIS fatigue short forms in rheumatoid arthritis. Qual Life Res 2018; 27: 2443–2451. DOI: 10.1007/s11136-018-1880-x.

20. Cook KF, Bamer AM, Roddey TS, et al. A PROMIS fatigue short form for use by individuals who have multiple sclerosis. Qual Life Res 2012; 21: 1021–1030. DOI: 10.1007/s11136-011-0011-8.

21. Bingham CO III, Gutierrez AK, Butanis A, et al. PROMIS fatigue short forms are reliable and valid in adults with rheumatoid arthritis. J Patient-Rep Outcomes 2019; 3. DOI: 10.1186/s41687-019-00105-6.

22. Carlozzi NE, Ianni PA, Tulsky DS, et al. Understanding health-related quality of life in caregivers of civilians and service members/veterans with traumatic brain injury: establishing the reliability and validity of PROMIS fatigue and sleep disturbance item banks. Arch Phys Med Rehabil 2019; 100: S102–s109. DOI: 10.1016/j.apmr.2018.05.020.

23. Cessna JM, Jim HSL, Sutton SK, et al. Evaluation of the psychometric properties of the PROMIS Cancer Fatigue Short Form with cancer patients. J Psychosomatic Research 2016; 81: 9–13. DOI: 10.1016/j.jpsychiores.2015.12.002.

24. Christodoulou C, Schneider S, Junghaenel DU, et al. Measuring fatigue directly using a brief scale adapted from the Patient-Reported Outcomes Measurement Information System (PROMIS). Qual Life Res 2014; 23: 1245–1253. DOI: 10.1007/s11136-013-0553-z.

25. Gibbons LE, Fredericksen R, Bates D, et al. Validity assessment of the PROMIS fatigue domain among people living with HIV. AIDS Research Therapy 2017; 14: 21. DOI: 10.1186/s12981-017-0146-y.

26. Hackney AJ, Klinedinst NJ and Resnick B. Measuring fatigue in older adults with joint pain: reliability and validity testing of the PROMIS fatigue short forms. J Nurs Meas 2019; 27: 534–553. DOI: 10.1891/1061-3749.27.3.534.

27. Hildenbrand AK, Quinn CT, Mara CA, et al. A preliminary investigation of the psychometric properties of PROMIS scales in emerging adults with sickle cell disease. Health Psychol 2019; 38: 386–390. DOI: 10.1037/hea0000696.

28. Kratz AL, Schilling S, Goelsing J, et al. The PROMIS FatigueFM Profile: a self-report measure of fatigue for use in fibromyalgia. Qual Life Res 2016; 25: 1803–1813. DOI: 10.1007/s11136-016-1230-9.

29. Pokrzywinski R, Soliman AM, Surrey E, et al. Psychometric assessment of the PROMIS Fatigue Short Form 6a in women with moderate-to-severe endometriosis-associated pain. J Patient-Reported Outcomes 2020; 4: 86–2010. DOI: 10.1186/s41687-020-00257-y/10/28.

30. Stone AA, Broderick JE, Junghaenel DU, et al. PROMIS fatigue, pain intensity, pain interference, pain behavior, physical function, depression, anxiety, and anger scales demonstrate ecological validity. J Clin Epidemiol 2016; 74: 194–206. DOI: 10.1016/j.jclinepi.2015.08.029.

31. Tomasson G, Farrar JT, Cuthbertson D, et al. Feasibility and construct validation of the patient reported outcomes measurement system information in systemic vasculitis. J Rheumatol 2019; 46: 928–934. DOI: 10.3899/jrheum.171405.

32. Yang M, Keller S and Lin J-MS. Psychometric properties of the PROMIS Fatigue Short Form 7a among adults with myalgic encephalomyelitis/chronic fatigue syndrome. Qual Life Res 2019; 28: 3375–3384. DOI: 10.1007/s11136-019-02289-4.

33. Yost KJ, Waller NG, Lee MK, et al. The PROMIS fatigue item bank has good measurement properties in patients with fibromyalgia and severe fatigue. Qual Life Res 2017; 26: 1417–1426. DOI: 10.1007/s11136-017-1501-0.

34. Yount SE, Atwood C, Donohue J, et al. Responsiveness of PROMIS to change in chronic obstructive pulmonary disease. J Patient-Reported Outcomes 2019; 3: 65. DOI: 10.1186/s41687-019-00255-4.

35. Cella D, Lai J-S, Jensen SE, et al. PROMIS fatigue item bank had clinical validity across diverse chronic conditions. J Clin Epidemiol 2016; 73: 128–134. DOI: 10.1016/j.jclinepi.2015.08.037.

36. Purvis TE, Neuman BJ, Riley LH 3rd et al. Physical function domain in spine patients. Spine 2017; 42: 921–929.

37. Egerton T, Ripphagen II, Nygård AJ, et al. Systematic content evaluation and review of measurement properties of questionnaires for measuring self-reported fatigue among older people. Qual Life Res 2015; 24: 2239–2255. DOI: 10.1007/s11136-015-0963-1.

38. Brodke DS, Goz V, Voss MW, et al. PROMIS PF CAT outperforms the ODI and SF-36 physical function domain in spine patients. Spine 2017; 42: 921–929. DOI: 10.1097/brs.000000000001965.
39. Fries J, Rose M and Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol 2011; 38: 1759–1764. DOI: 10.3899/jrheum.110402.

40. Elsman EBM, Roorda LD, Crins MHP, et al. Dutch reference values for the patient-reported outcomes measurement information system scale v1.2 - global health (PROMIS-GH). J Patient-Reported Outcomes 2021; 5. DOI: 10.1186/s41687-021-00314-0.

41. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?ts=1649919922005

42. Hortensius L. Advanced Measurement – Logistic Regression for DIF Analysis. Minneapolis, M.N.: University of Minnesota, 2012.

43. Choi SW, Gibbons LE and Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Statistical Software 2011; 39: 1–30.

44. Deveillis R. PROMIS 1 Social Supplement, 2016, https://dataverse.harvard.edu/dataverse.xhtml?alias=HealthMeasures.

45. Terwee CB, Crins MHP, Roorda LD, et al. International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. J Clin Epidemiol 2021; 134: 1–13. DOI: 10.1016/j.jclinepi.2021.01.011.

46. http://www.healthmeasures.net/promis.

47. Cappelleri JC, Jason Lundy J and Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014; 36: 648–662. DOI: 10.1016/j.clinthera.2014.04.006.

48. Terwee CB, Crins MHP, Roorda LD, et al. International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. J Clin Epidemiol 2021; 134: 1–13. DOI: 10.1016/j.jclinepi.2021.01.011.

49. NIIPHat Environment. RIVM forecasting study: a healthier Netherlands with more people living with a chronic disease, https://www.rivm.nl/en/news/rivm-forecasting-study-a-healthier-netherlands-with-more-people-living-with-a-chronic-disease#:~:text=One%20of%20the%20most%20important,morbidity%20will%20also%20grow.

50. Calvert M, Kyte D, Price G, et al. Maximising the impact of patient reported outcome assessment for patients and society. BMJ 2019; 364: k5267.

51. Jim HSL, Hoogland AI, Brownstein NC, et al. Innovations in research and clinical care using patient-generated health data. CA: Cancer J Clin 2020; 70: 182–199.

52. Eton DT, Beebe T, Hagen P, et al. Harmonizing and consolidating the measurement of patient-reported information at health care institutions: a position statement of the Mayo Clinic. Patient Relat Outcome Measures 2014; 5: 7.

53. Noonan VK, Cook KF, Bamer AM, et al. Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form. Qual Life Res 2012; 21: 1123–1133. DOI: 10.1007/s11136-011-0040-3.

54. Stone PROseta. http://www.prosettastone.org/Pages/default.aspx, accessed 3 11 2019.

55. Rose M, Bjorner JB, Gandek B, et al. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014; 67: 516–526. DOI: 10.1016/j.jclinepi.2013.10.024.

56. Fries JF, Cella D, Rose M, et al. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009; 36: 2061–2066. DOI: 10.3899/jrheum.090358.

57. Gausden EB, Levack AE, Sin DN, et al. Validating the Patient Reported Outcomes Measurement Information System (PROMIS) computerized adaptive tests for upper extremity fracture care. J Shoulder Elbow Surg 2018; 27: 1191–1197. DOI: 10.1016/j.jse.2018.01.014.

58. Hung M, Stuart AR, Higgins TF, et al. Computerized adaptive testing using the PROMIS physical function item bank reduces test burden with less ceiling effects compared with the short musculoskeletal function assessment in orthopaedic trauma patients. J Orthopaedic Trauma 2014; 28: 439–443. DOI: 10.1097/bot.0000000000000059.

59. Crins MHP, van der Wees PJ, Klausch T, et al. Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. PLoS One 2018; 13: e0192187. DOI: 10.1371/journal.pone.0192187.

60. Luijten MAJ, van Litsenburg RRL, Terwee CB, et al. Psychometric properties of the Patient-Reported Outcomes Measurement Information System (PROMIS) pediatric item bank peer relationships in the Dutch general population. Qual Life Res 2021; 30: 2061–2070. DOI: 10.1007/s11136-021-02781-w.

61. Terwee CB, Crins MHP, Boers M, et al. Validation of two PROMIS item banks for measuring social participation in the Dutch general population. Qual Life Res 2019; 28: 211–220. DOI: 10.1007/s11136-018-1995-0.

62. Flens G, Smits N, Terwee CB, et al. Development of a computerized adaptive test for anxiety based on the Dutch-Flemish version of the PROMIS item bank. Assessment 2019; 26: 1362–1374. DOI: 10.1177/1073191117746742.

63. Katz P, Pedro S, Alemao E, et al. Estimates of responsiveness, minimally important differences, and patient acceptable symptom state in five patient-reported outcomes measurement information system short forms in systemic lupus erythematosus. ACR Open Rheumatology 2020; 2: 53–60. DOI: 10.1002/acr2.11100.

64. Lapin B, Thompson NR, Schuster A, et al. Clinical utility of patient-reported outcome measurement information system
domain scales. *Circ Cardiovasc Qual Outcomes* 2019; 12: e004753. DOI: 10.1161/circoutcomes.118.004753.

65. Yost KJ, Eton DT, Garcia SF, et al. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *J Clin Epidemiol* 2011; 64: 507–516. DOI: 10.1016/j.jclinepi.2010.11.018.

66. Kasturi S, Szymonifka J, Burket JC, et al. Validity and reliability of patient reported outcomes measurement information system computerized adaptive tests in systemic lupus erythematosus. *J Rheumatology* 2017; 44: 1024–1031. DOI: 10.3899/jrheum.161202.

67. Hu Lt and Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equation Model A Multidisciplinary J* 1999; 6: 1–55.

68. Rodriguez A, Reise SP and Haviland MG. Applying bifactor statistical indices in the evaluation of psychological measures. *J Personal Assess* 2016; 98: 223–237. DOI: 10.1080/00223891.2015.1089249.

69. Reise SP, Scheines R, Widaman KF, et al. Multidimensionality and structural coefficient bias in structural equation modeling. *Educ Psychol Meas* 2013; 73: 5–26. DOI: 10.1177/0016491412449831.

70. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks. *Med Care* 2007; 45: S22–S31. DOI: 10.1097/01.mlr.0000250483.85507.04.

71. Orlando M and Thissen D. Further investigation of the performance of S - X2: an item fit index for use with dichotomous item response theory models. *Appl Psychol Meas* 2003; 27: 289–298.

72. McKinley RL and Mills CN. A comparison of several goodness-of-fit statistics. *Appl Psychol Meas* 1985; 9: 49–57.