An initial psychometric evaluation of the German PROMIS® v1.2 Physical Function item bank in patients with a wide range of health conditions

Gregor Liegl, Matthias Rose, Helena Correia, Felix Fischer, Sibel Kanlidere, Annett Mierke, Alexander Obbarius, Sandra Nolte

Citation details
Liegl G, Rose M, Correia H, Fischer HF, Kanlidere S, Mierke A, et al. An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions. Clinical Rehabilitation. [Online] SAGE Publications; 2017;32(1): 84–93. DOI: 10.1177/0269215517714297

Clinical Rehabilitation

An initial psychometric evaluation of the German PROMIS® v1.2 Physical Function item bank in patients with a wide range of health conditions

Journal:	*Clinical Rehabilitation*
Manuscript ID	CRE-2016-5658.R2
Manuscript Type:	Original Article
Keywords:	Physical Function, Patient-reported Outcomes, Cultural Adaptation, PROMIS, Nonparametric Item Response Theory

Abstract:

Objectives: To translate the PROMIS® Physical Function (PF) item bank version 1.2 into German, and to investigate psychometric properties of resulting full bank and seven derived short forms.
Design: Cross-sectional psychometric study.
Setting: Inpatient and outpatient clinics of the Department of Psychosomatic Medicine at Charité - Universitätsmedizin Berlin, Germany.
Subjects: Ten adult patients with various chronic diseases participated in cognitive debriefing interviews. The final item bank was administered to n=266 adult patients with a broad range of medical conditions.
Interventions: Patient-reported outcome assessment as part of routine care.
Main measures: PROMIS v1.2 PF bank; MOS SF-36® PF scale (PF-10).
Results: Cross-cultural adaptation of the item bank followed established guidelines. For the final German translation, the corrected item-total correlations ranged from 0.44 to 0.84. Cronbach's Alpha was high for each PROMIS PF short form (α=0.88-0.96). The full PROMIS PF bank and most short forms correlated highly with the SF-36 PF-10 (r=0.85-0.90), with the exception of PROMIS Upper Extremity (r=0.64). PROMIS Upper Extremity showed ceiling effects and lower agreement with the full bank than other short forms. Unidimensionality was supported for all PROMIS PF measures using traditional factor analysis and nonparametric item response theory.
Conclusions: The German PROMIS PF bank was found to be conceptually equivalent to the English version and fulfilled the psychometric requirements for use of short forms in clinical practice. Future studies should pay particular attention to samples with upper extremity functional limitations to further investigate the dimensional structure of physical function as conceptualized according to PROMIS.

An initial psychometric evaluation of the German PROMIS® v1.2 Physical Function item bank in patients with a wide range of health conditions

Gregor Liegl[1*], Matthias Rose[1,2], Helena Correia[3], Felix Fischer[1,4], Sibel Kanlidere[1], Annett Mierke[1], Alexander Obbarius[1], Sandra Nolte[1,5]

[1] Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Berlin, Germany
[2] Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA
[3] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
[4] Institute for Social Medicine, Epidemiology and Health Economics, Charité Universitätsmedizin Berlin, Germany
[5] Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Burwood, Australia

[*]Correspondence address: Gregor Liegl, Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany, Tel.: +49 30 450 565259, email: gregor.liegl@charite.de, institutional URL: http://psychosomatik.charite.de/

Abstract

Objectives: To translate the PROMIS® Physical Function (PF) item bank version 1.2 into German, and to investigate psychometric properties of resulting full bank and seven derived short forms.

Design: Cross-sectional psychometric study.

Setting: Inpatient and outpatient clinics of the Department of Psychosomatic Medicine at Charité - Universitätsmedizin Berlin, Germany.

Subjects: Ten adult patients with various chronic diseases participated in cognitive debriefing interviews. The final item bank was administered to n=266 adult patients with a broad range of medical conditions.

Interventions: Patient-reported outcome assessment as part of routine care.

Main measures: PROMIS v1.2 PF bank; MOS SF-36® PF scale (PF-10).

Results: Cross-cultural adaptation of the item bank followed established guidelines. For the final German translation, the corrected item-total correlations ranged from 0.44 to 0.84. Cronbach's Alpha was high for each PROMIS PF short form (α=0.88-0.96). The full PROMIS PF bank and most short forms correlated highly with the SF-36 PF-10 (r=0.85-0.90), with the exception of PROMIS Upper Extremity (r=0.64). PROMIS Upper Extremity showed ceiling effects and lower agreement with the full bank than other short forms. Unidimensionality was supported for all PROMIS PF measures using traditional factor analysis and nonparametric item response theory.

Conclusions: The German PROMIS PF bank was found to be conceptually equivalent to the English version and fulfilled the psychometric requirements for use of short forms in clinical practice. Future studies should pay particular attention to samples with upper extremity functional limitations to further investigate the dimensional structure of physical function as conceptualized according to PROMIS.

Introduction

Self-rated physical function (PF) is a core patient-reported outcome (PRO) in clinical trials evaluating the effectiveness of therapy and rehabilitation programs for various chronic diseases.[1-3] In clinical practice, routine collection of patient-reported PF can be useful to optimize treatment plans and to improve the communication between patients and clinicians.[4, 5] Many different PROs assessing PF have been developed so far.[6] However, the specific scores of different instruments are often not scaled on the same metric, affecting the comparability across different medical fields, diseases, and interventions.[7] To overcome these potential limitations of comparability, item response theory (IRT) can be used for the development of generic item banks by calibrating any number of items that are aimed to measure the same latent construct on a common metric.[8, 9] The Patient-Reported Outcomes Measurement Information System (PROMIS®) initiative[10], funded by the U.S. National Institutes of Health (NIH), is one of the most extensive projects providing IRT-based item banks for construct-based assessment of many domains of health-related quality of life[11], including a comprehensive item bank measuring physical function.[12] The psychometric properties of the PROMIS PF item bank have been evaluated in several clinical and non-clinical populations.[13, 14]

The translation and cross-cultural adaptation of the PROMIS PF items for use in non-English-speaking populations is an important effort to ensure comparability between samples with different languages and cultural backgrounds. Recent findings of psychometric studies investigating differential item functioning (DIF) by language are inconsistent. For instance, while the Dutch version of the PROMIS PF item bank yields PF scores that are largely comparable to the U.S. version[15], Spanish-speaking participants responded differently to almost half of all PF items compared with English-speaking participants when the same

underlying level of functioning was assumed.[16] These findings emphasize the need for separate psychometric evaluations of each language version.

In this paper, we describe the translation and initial psychometric evaluation of the German version of the PROMIS v1.2 Physical Function item bank for use in adult patients with various medical conditions.

Methods

The PROMIS® Physical Function item bank version 1.2

The PROMIS PF bank was developed as a generic item bank for the assessment of physical function in various clinical and non-clinical populations.[7, 12] Version 1.2 *(PROMIS Bank v1.2 – Physical Function*) includes 121 items, which can be administered as short forms or as computerized adaptive tests (CAT).[17] In this study we evaluated the psychometric properties of the full bank and seven PROMIS PF short forms: Mobility (15 items), Upper Extremity (16 items), and five generic short forms of different lengths: SF-4a, SF-6b, SF-8b, SF-10a, SF-20a (4, 6, 8, 10, and 20 items, respectively) (www.assessmentcenter.net).

German translation and cognitive debriefing

The PROMIS v1.2 PF bank was translated into German by a bilingual expert group, according to the Functional Assessment of Chronic Illness Therapy (FACIT) translation methodology[18], which includes forward and back translations, reviews from different German-speaking countries, reconciliation meetings, quality review, and cognitive debriefing interviews (for detailed information on the translation core steps, see Appendix 1). To test understandability and clarity of the translated items, cognitive debriefing interviews were conducted with ten patients of the Department of Rheumatology and the Department of Psychosomatic Medicine at Charité – Universitätsmedizin Berlin between June and July 2014.

Evaluating psychometric properties

Data collection

The full German PROMIS v1.2 PF item bank was administered to a clinically diverse sample of adult patients of the Department of Psychosomatic Medicine at Charité. This included patients with somatoform disorders, chronic pain, and eating disorders, but also patients with a variety of physical conditions associated with mental disorders and psychological distress (for example, major depression following cancer diagnosis or exacerbated diabetes mellitus due to a depressive episode). The majority of data were collected consecutively as part of routine PRO assessment at the outpatient clinic of the Department between June 2015 and February 2016. These patients completed the PROMIS PF item bank and other questionnaires by themselves electronically using personal digital assistant (PDA) devices, which were handed out in the clinic's waiting room. In addition, a small proportion of participants answered paper-based questionnaires at the inpatient clinic of the Department of Psychosomatic Medicine between August and September 2015. Inclusion criteria were age ≥18 years and German language fluency.

Data preparation and scoring

For the full PROMIS PF bank and each of the short forms we calculated scale sum scores following the PROMIS PF Scoring Manual, with higher scores indicating higher function. To enable direct comparisons between the different PF measures (full bank and short forms), we standardized respective scores using z-score transformation (mean=0.0; SD=1.0).

Psychometric analyses

Psychometric properties were evaluated following frequently used criteria in patient-reported outcomes.[19] Ceiling and floor effects were considered to be present if more than 15% of participants achieved the maximum or minimum scale score, respectively.[19] Items

with more than 95% of responses in one category were considered to be insufficient for psychometric evaluation.[7]

Internal consistency of each short form was evaluated using Cronbach's Alpha.[19] An alpha value below 0.70 was considered to be insufficient, between 0.80 and 0.90 as high and above 0.90 as very high. The corrected item-total correlations were calculated for each item of the full item bank as an indicator of discriminative power (r_{itc}, cut-off >0.40).

Construct validity was evaluated using Pearson correlations between PROMIS PF measures and the 10-item MOS SF-36 physical functioning scale (SF-36 PF-10).[19, 20] Moreover, we calculated the correlations between the full PROMIS PF bank and each short form and investigated the extent of scoring discrepancy by calculating root mean square errors (RMSEs) for respective z-scores.

To evaluate unidimensionality of the full PROMIS PF bank as well as the short forms, we used the monotone homogeneity model (MHM), a nonparametric item response theory (NIRT) approach that has been recommended for PROs with polytomous items.[21, 22] Model fit was investigated using Loevinger's homogeneity coefficient *H*, with *H*>0.5 indicating a strong unidimensional scale. Item-specific H_j coefficients determine discriminative power of each item for a given scale, with H_j>0.3 indicating sufficient contribution to the measurement.[21] Significant violations of the monotonicity assumption were checked for each item.[23]

For the Mobility scale, the Upper Extremity scale, and short forms SF-8b (covering all items included in SF-4a and SF-6b) and SF-20a (covering all items included in SF-10a), we additionally conducted confirmatory factor analyses (CFA) using the diagonally weighted least squares (DWLS) estimator as recommended for ordinal data in the PROMIS Scientific Standards document (www.nihpromis.org/Documents/PROMIS_Standards_050212.pdf).

Model fit was evaluated by calculating chi-square statistics, the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and 90% confidence intervals of the Root Mean Square Error of Approximation (RMSEA). CFI and TLI values of 0.95 or larger were considered to indicate good model fit, while RMSEA values of 0.08 or smaller were considered to indicate unidimensionality. Items with factor loadings >0.70 were considered to be significant contributors to the latent trait. A residual correlation <0.25 for each pair of items was considered to indicate locale independence.[12] For the short forms, we additionally fitted exploratory bifactor models with one general factor (representing the latent trait being assessed, i.e. physical functioning) and allowed three exploratory group factors.[24] An explained common variance (ECV) by the general factor of 0.60 or higher was considered as an indicator of sufficient unidimensionality of respective model.[24, 25]

For all statistical analyses, *R 3.0.1* was applied using the packages *lavaan*, *mokken*, and *psych.*[26-29]

Ethics approval

The study was approved by the Charité Ethics Committee, number EA1/119/15.

Results

Translation and cultural adaptation

Two major decisions were made during the translation process: First, the English-language item stem 'Are you able to' was translated with 'Können Sie' (which literally means 'Can you') because in German these expressions are semantically similar but the latter is more straightforward. Second, as 'door knobs' are largely uncommon in German-speaking countries (Austria, Germany, Switzerland), we used the German word for 'door handles'

instead. During the cognitive debriefing interviews, some additional expressions were challenged by the participants. While the translated version of the item bank was generally well understood by most participants, the chosen German translation for 'walking up one/several/five flight(s) of stairs' ('Treppenabsatz') was criticized by all patients and was therefore replaced by the German wording for 'walking up one/several/five floor(s)'.

Further information on the results of the extensive translation process are presented in Appendix 2.

Psychometric properties of the German PROMIS PF item bank

Sample

Data were collected from 266 patients of the Department of Psychosomatic Medicine with various medical conditions (89.8% inpatients using personal digital assistant devices, 10.2% outpatients using paper-based questionnaires). Clinician-reported diagnoses according to the diagnostic criteria of the 10th revision of the International Statistical Classification of Diseases (ICD-10)[30] and other patient characteristics are presented in Table 1. The average age of participants was 43 years (SD=15), n=163 were female (61%). The mean physical function level as measured by the SF-36 PF-10 sum score was 22.5 (SD=6.0), which is about 0.7 standard deviations lower than the mean of the German general population.[31]

**** Table 1 about here ****

Distribution of PROMIS PF data

Scale characteristics for the full item bank and each short form are presented in Table 2. Among all short forms, highest skewness was found for the Upper Extremity scale scores

with a value of -2.56. Skewness was less than or equal to an absolute value of 1.20 for all other short forms and for the full item bank. Further, in the Upper Extremity scale, 38% of the 266 participants reached the highest possible scale score, indicating ceiling effects. Ceiling effects were also found for PROMIS SF-4a.

Table A.1 (Appendix 3) shows the individual item characteristics of those 50 PROMIS PF items that are included in at least one of the PF short forms (item characteristics of the remaining 71 items not shown). No individual item included in the full item bank had more than 95% of responses in one category. A total of seven items in the full item bank were highly skewed with a value <-4, all of which were related to hand function (i.e., grip or fine motor activities; data partially not shown).

**** Tables 2 about here ****

Internal Consistency

The internal consistency was high or very high for all PROMIS PF short forms (Table 2). For each individual item in the full item bank, the corrected item-total correlations was >0.40 (range: 0.44 to 0.84). Compared to other PF subdomains, item discrimination tended to be lowest for upper extremity items, especially when asking about fine motor skills (e.g., items A20, B21, or A35; see Appendix 3: Table A.1).

Construct validity

The correlation with the SF-36 PF-10 was high for the full PROMIS PF item bank, the generic PROMIS PF short forms (SF-4a, SF-6b, SF-8b, SF-10a, and SF-20a), and the Mobility scale (r=0.85 to 0.90; Table 2). In contrast, the correlation between the SF-36 PF-10 and the

PROMIS Upper Extremity scale was considerably lower (r=0.64). All PROMIS PF short forms correlated highly with the full item bank (r≥0.87).

The average scoring discrepancy with the full PROMIS PF item bank was highest for the Upper Extremity scale (RMSE=0.54), despite having a relatively high number of items compared to most other short forms (Table 2).

Unidimensionality

Using nonparametric item response theory (NIRT), unidimensional model fit was supported (Table 2). The scaling coefficient *H* exceeded the threshold for a strong unidimensional scale in the full item bank (*H*=0.646) and in each short form (*H*=0.595 to 0.743). H_j coefficients of all items were considerably higher than 0.3 (H_j=0.439 to 0.838); lowest values were found for hand function items. We did not find violations of monotonicity for any item in the full bank.

Results of confirmatory factor analyses (CFA) and bifactor analyses are presented in Table 3. In the CFAs, we found statistically significant Chi-square values in all short forms. However, for each scale both the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) were above 0.99 and the lower bound of the 90% confidence interval of the Root Mean Square Error of Approximation (RMSEA) did not exceed 0.08. All residual correlations were lower than 0.25 and all factor loadings were higher than 0.70. Exploratory bifactor analyses resulted in high values of explained common variance of the general factor (ECV=0.64 to 0.80), indicating sufficient unidimensionality.

**** Table 3 about here ****

Discussion

All 121 items of the German translation of the PROMIS v1.2 Physical Function item bank, as part of the full item bank and also as part of the short forms, showed satisfactory psychometric properties. Hence, our results support clinical use of the translated PROMIS PF measures in German-speaking adult patient populations.

During the translation and psychometric evaluation we had to overcome several challenges that were mainly due to the broadly defined PROMIS PF construct covered by the considerable number of 121 items.[7] First, after extensive discussions within the bilingual expert group, a total of five items were translated into German with somewhat modified content. Interestingly, some of the changes made were similar to those made in the Dutch translation[32] and other languages.[33] For example, 'door knobs' appear to be unusual in the Netherlands as well, with the final translation also using the expression 'door handles'. While this supports our choice of wording for cultural adaptation, the specific hand movements required to open a door differ between using a knob versus using a handle, i.e. turning versus pressing down. Therefore, there is a potential threat that the culturally adapted item will show different measurement characteristics compared to the original English-language item. Similarly, our decision to use 'floors' instead of 'flights of stairs' potentially makes these items more difficult in German compared to the English original. It remains to be shown if the relative position of these items on the PF continuum is equal to the original English-language item bank calibrated in the U.S. population.[12] As a result, future studies will have to investigate whether the items included in the German item bank, especially those with modified content, show language-related differential item functioning (DIF). The detection of (and accounting for) DIF is an important requirement for cross-cultural comparability of scores.[15, 16]

Second, verifying sufficient unidimensional fit is an important prerequisite for the specific item calibration approach as conducted by PROMIS.[9] However, for the special case of PRO item banks, it has been shown that common fit indices were negatively affected by the large number of included items.[25] Moreover, using traditional factor analysis to investigate the dimensional structure of as many as 121 items requires very large samples, especially when using approaches as recommended for ordinal data and when data are skewed.[34, 35] For reasons of efficiency, it is highly preferable to detect potentially misfitting items prior to conducting expensive calibration studies. Therefore, as an efficient way to explore the unidimensional fit in an early stage, we applied a nonparametric item response theory model: the monotone homogeneity model (MHM).[22] For highly discriminating items, robust MHM results have been found for a sample size of n=250, independent from the number of included items.[36] The MHM is a general case of the more commonly used parametric graded response model (GRM), which has been suggested by PROMIS for item bank calibration.[10] Thus, if misfit is identified for the MHM, it can be concluded that the more restrictive assumptions of the GRM are violated as well. In the present study, the MHM analyses identified the German PROMIS PF item bank (and each related short form) to be a strong unidimensional scale, suggesting that no further adjustments have to be made before carrying out further validation and calibration studies in larger samples.

A third challenge is related to the interpretation of the results found for the upper extremity items. Although a considerable below-average level of physical function was identified for the study sample, a high number of participants showed ceiling effects in the Upper Extremity scale. Moreover, Upper Extremity scores were only moderately correlated to the external SF-36 PF-10 criterion and showed highest scoring discrepancies with the full PROMIS PF bank. These findings suggest a somewhat different underlying construct of PF for

upper extremity items compared to other PF domains. It is also noteworthy that especially those items related to hand activities showed lower correlations to the remaining items in the item bank. Both the presence of ceiling effects in upper extremity items and potential problems of using one common construct for the assessment of different subdomains of PF have been previously reported for the English-language PROMIS PF item bank.[37] Thus, although good psychometric properties for all German PROMIS PF items were verified in this initial psychometric evaluation, further validation studies are necessary and should include a sufficiently large number of patients from different clinical subgroups, especially including those with impaired hand function, to further investigate dimensionality aspects and examine potential differential item functioning by disease.

Despite rigorous methods used for both cultural adaptation and psychometric evaluation, this study has some limitations. First, although item response theory based item parameters are available for the English-language PROMIS PF item bank, allowing for computer adaptive testing[17], it is not clear if all of these parameters are transferable to the German population. To verify whether U.S. parameters can be used in German language, analysis of differential item functioning need to be conducted in future studies. Second, our results are based on a rather young sample consisting of inpatients and outpatients of only one medical clinic. It is possible that different test and item characteristics would have been identified in other patient populations. Therefore, further studies including other disease groups will be important to confirm our results of this initial psychometric evaluation. A third potential limitation is that we used different methods for data collection. While the vast majority of participants responded to the item bank electronically, we used paper-based questionnaires in some cases. However, for U.S. samples it has been shown that the method of administration of PROMIS PF items did not bias resulting score levels.[38]

Finally, for the English PROMIS PF bank, version 2.0 has recently been launched, including 44 additional items (next to the items of version 1.2), which have not been translated into German yet. However, while in version 2.0 the Upper Extremity scale has been expanded (referred to as “Upper Extremity v2.0 item bank”), the content of the Mobility scale and the five generic short forms remained unchanged, compared to version 1.2. Moreover, once the PROMIS v2.0 PF bank will be presented for German-speaking populations in the future, scores across versions 1.2 and 2.0 will be directly comparable to each other as the underlying PROMIS PF metric remains unchanged when calibrating new items to the item bank.

In conclusion, the translation and cultural adaptation of the PROMIS v1.2 PF item bank was successful. We found strong evidence that the German version of the item bank is conceptually equivalent to the original version. The German PROMIS v1.2 PF item bank demonstrated good measurement properties, comparable to those found for the English item bank and other language versions. The psychometric properties of seven PROMIS PF short forms were found to be satisfactory for clinical use in German-speaking patients. All German PROMIS PF short forms validated in this work are available upon request from the authors.

Clinical Messages

- The German translation of the PROMIS v1.2 Physical Function item bank was found to be conceptually equivalent to the original English version.
- The psychometric properties of the German PROMIS PF bank and derived short forms were found to be satisfactory for clinical use.
- All 121 PROMIS PF items fulfilled the requirements for further calibration studies in large German-speaking samples using parametric item response theory.

Author contributions: All authors have made substantial contributions, and have read, revised, and approved the paper. SN, GL, and MR conceptualized and designed the study. SN, MR, FF, SK, HC, and GL participated in the German translation of the PROMIS PF item bank. Cognitive debriefing interviews were carried out by SK, while data acquisition for the psychometric evaluation was organized and performed by AM, AO, and GL. Analysis and interpretation of data was carried out by GL, SN, FF, and HC. GL was responsible for writing and preparing the manuscript.

Conflicts of interest: None to declare.

Funding: This research was supported by a Rahel-Hirsch scholarship from the Charité - Universitätsmedizin Berlin to SN, covering both the translation and qualitative testing of the PROMIS v1.2 PF item bank.

References

1. Kluetz PG, Slagle A, Papadopoulos E, et al. Focusing on Core Patient-Reported Outcomes in Cancer Clinical Trials: Symptomatic Adverse Events, Physical Function, and Disease-Related Symptoms. *Clin Cancer Res*. 2016; 22: 1553-8.
2. Artz N, Dixon S, Wylde V, et al. Comparison of group-based outpatient physiotherapy with usual care after total knee replacement: A feasibility study for a randomized controlled trial. *Clin Rehabil*. 2016; 31: 487-99.
3. van Tuyl LHD and Boers M. Patient-reported outcomes in core domain sets for rheumatic diseases. *Nat Rev Rheumatol*. 2015; 11: 705-12.
4. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care*. 2007; 45: S3.
5. Snyder CF, Herman JM, White SM, et al. When using patient-reported outcomes in clinical practice, the measure matters: a randomized controlled trial. *J Oncol Pract*. 2014; 10: e299-e306.
6. Bruce B, Fries JF, Ambrosini D, et al. Better assessment of physical function: item improvement is neglected but essential. *Arthritis Res Ther*. 2009; 11: R191.
7. Rose M, Bjorner JB, Becker J, Fries JF and Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol*. 2008; 61: 17-33.
8. Liegl G, Wahl I, Berghofer A, et al. Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. *J Clin Epidemiol*. 2016; 71: 25-34.
9. Cella D, Gershon R, Lai J-S and Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007; 16: 133-41.
10. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. 2010; 63: 1179-94.
11. Ader DN. Developing the patient-reported outcomes measurement information system (PROMIS). *Med Care*. 2007; 45: S1-S2.
12. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF and Ware JE, Jr. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol*. 2014; 67: 516-26.
13. Oude Voshaar MA, Ten Klooster PM, Glas CA, Vonkeman HE, Krishnan E and van de Laar MA. Relative performance of commonly used physical function questionnaires in rheumatoid arthritis and a patient-reported outcomes measurement information system computerized adaptive test. *Arthritis Rheumatol*. 2014; 66: 2900-8.
14. Liegl G, Gandek B, Fischer HF, et al. Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. *Arthritis Res Ther*. 2017; 19: 66.
15. Oude Voshaar MA, ten Klooster PM, Glas CA, et al. Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. *PLoS One*. 2014; 9: e92367.
16. Paz SH, Spritzer KL, Morales LS and Hays RD. Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Qual Life Res*. 2012; 22: 1819-30.
17. Bjorner JB, Chang C-H, Thissen D and Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. *Qual Life Res*. 2007; 16: 95-108.
18. Eremenco SL, Cella D and Arnold BJ. A Comprehensive Method for the Translation and Cross-Cultural Validation of Health Status Questionnaires. *Eval Health Prof*. 2005; 28: 212-32.
19. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007; 60: 34-42.

20. Ware Jr JE and Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Med Care*. 1992: 473-83.
21. Stochl J, Jones PB and Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Med Res Methodol*. 2012; 12: 1-16.
22. Sijtsma K, Emons WH, Bouwmeester S, Nyklicek I and Roorda LD. Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). *Qual Life Res*. 2008; 17: 275-90.
23. Sijtsma K and Molenaar IW. *Introduction to nonparametric item response theory*. London: Sage Publications, 2002.
24. Reise SP, Scheines R, Widaman KF and Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educ Psychol Meas*. 2013; 73: 5-26.
25. Cook KF, Kallen MA and Amtmann D. Having a Fit: Impact of Number of Items and Distribution of Data on Traditional Criteria for Assessing IRT's Unidimensionality Assumption. *Qual Life Res*. 2009; 18: 447-60.
26. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2012.
27. Rosseel Y. lavaan: An R package for structural equation modeling. *J Stat Softw*. 2012; 48: 1-36.
28. Van der Ark LA. Mokken scale analysis in R. *J Stat Softw*. 2007; 20: 1-19.
29. Revelle W. *psych: Procedures for psychological, psychometric, and personality research*. Illinois: Northwestern University, 2014.
30. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems. 10th Revision*. Geneva: World Health Organization, 2010.
31. Ellert U and Kurth B-M. Health-related quality of life in adults in Germany. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*. 2013; 56: 643-9.
32. Oude Voshaar MAH, ten Klooster PM, Taal E, Krishnan E and van de Laar MAFJ. Dutch translation and cross-cultural adaptation of the PROMIS® physical function item bank and cognitive pre-test in Dutch arthritis patients. *Arthritis Res Ther*. 2012; 14: R47.
33. Parks-Vernizzi E, Correia H, Arnold B and Cella D. Promis® Physical Function Multilingual Translations: Progress And Outlook. *Value Health*. 2015; 18: A744-A5.
34. Mindrila D. Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*. 2010; 1: 60-6.
35. Hu L and Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychol Methods*. 1998; 3: 424-53.
36. Straat JH, van der Ark LA and Sijtsma K. Minimum sample size requirements for Mokken scale analysis. *Educ Psychol Meas*. 2014; 74: 809-22.
37. Hung M, Clegg DO, Greene T and Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res*. 2011; 29: 947-53.
38. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU and Ware JE. Method of Administration of PROMIS Scales Did Not Significantly Impact Score Level, Reliability or Validity. *J Clin Epidemiol*. 2014; 67: 108-13.

Tables

Table 1: Demographic and clinical characteristics of the patients of the Department of Psychosomatic Medicine at Charité - Universitätsmedizin Berlin (n=266)

Mean age (SD)	42.7 (14.9)	
n female (%)	163 (61.3)	
n > secondary school certificate (%)	122 (45.9)	
Most frequent clinician-reported diagnoses (according to ICD-10 diagnostic criteria): *	*n*	*%*
F30–F39: Mood (affective) disorders	86	32.3
F45: Somatoform disorders	79	29.7
E00–E90: Endocrine, nutritional and metabolic diseases	59	22.2
F40-F42: Anxiety disorders or obsessive-compulsive disorder	51	19.2
F50–F59: Behavioural syndromes associated with physiological disturbances and physical factor	49	18.4
F43: Reaction to severe stress, and adjustment disorders	33	12.4
M00–M99: Diseases of the musculoskeletal system and connective tissue	30	11.3
K00–K93: Diseases of the digestive system	28	10.5
G00–G99: Diseases of the nervous system	27	10.2
I00–I99: Diseases of the circulatory system	27	10.2

Abbreviations: ICD-10, 10th revision of the International Statistical Classification of Diseases and Related Health Problems; SD, standard deviation;

* ICD-10 categories that were reported for more than 10% of the study sample (n=266); multiple diagnoses per participant are possible; for detailed information on the clinical diagnoses, see Appendix 3: Table A.1.

Table 2: Scale characteristics of PROMIS v1.2 PF measures

PROMIS PF measure	Scale score [a] Range	Scale score [a] Mean (SD)	Skew-ness	% ceiling [b]	Cronbach's Alpha	Correlation with Full Bank	Correlation with SF-36 PF-10	RMSE [c]	*H*
Full Bank (121 items)	221-605	509.5 (80.5)	-1.20	0.4	0.99 [d]	-	0.87	-	0.646
SF-4a (4 items)	4-20	16.2 (3.9)	-1.14	22.2	0.88	0.89	0.85	0.48	0.713
SF-6b (6 items)	6-30	22.9 (5.9)	-0.74	12.8	0.91	0.92	0.89	0.39	0.716
SF-8b (8 items)	8-40	29.6 (8.2)	-0.54	12.0	0.94	0.94	0.90	0.36	0.743
SF-10a (10 items)	14-50	39.6 (8.2)	-0.80	5.6	0.92	0.97	0.90	0.26	0.737
SF-20a (20 items)	37-100	83.0 (14.3)	-1.0	4.5	0.95	0.99	0.88	0.17	0.696
Mobility (15 items)	16-75	61.3 (13.8)	-1.16	11.3	0.96	0.96	0.87	0.28	0.692
Upper Extremity (16 items)	31-80	75.1 (8.1)	-2.56	38.0	0.94	0.87	0.64	0.54	0.595

Abbreviations: ECV, explained common variance; H, Loevinger's homogeneity coefficient; PF, physical function; RMSE, root mean square error; SD, standard deviation;

[a] Scale score = (raw sum score * number of items in respective scale)/(number of items that were actually answered)

[b] Percentage who reached the highest ('ceiling') possible scale score, based on the full sample (n=266)

[c] For calculating the root mean square error (RMSE) between full item bank and short forms, z-scores of respective scales were used

[d] Cronbach`s Alpha for the full bank is of little relevance for clinical practice, as usually not all 121 items are administered

Table 3: Results of the Factor Analyses of PROMIS v1.2 PF short forms

	Confirmatory factor analysis (one-factor model)					Exploratory bifactor model
PROMIS PF Short Form	**Chi² (df)**	**p**	**CFI [c]**	**TLI [d]**	**RMSEA (90% CI) [e]**	**ECV [f]**
SF-8b [a]	56.8 (20)	<.001	0.998	0.998	0.083 (0.058-0.109)	0.75
SF-20a [b]	456.4 (170)	.007	0.996	0.995	0.080 (0.071-0.089)	0.64
Mobility	242.2 (90)	<.001	0.997	0.997	0.080 (0.068-0.092)	0.80
Upper Extremity	171.0 (104)	<.001	0.998	0.997	0.049 (0.036-0.062)	0.73

Abbreviations: CFI, Comparative Fit Index; CI, confidence interval; df, degrees of freedom; ECV, explained common variance; PF, physical function; RMSEA, Root Mean Square Error of Approximation; SF, short form; TLI, Tucker-Lewis Index;

[a] The PROMIS PF SF-8b is covering all items of PROMIS PF SF-6b and PROMIS PF SF-4a.

[b] The PROMIS PF SF-20a is covering all items of PROMIS PF SF-10a.

[c] A CFI value of 0.95 or larger is considered to indicate good model fit.

[d] A TLI value of 0.95 or larger is considered to indicate good model fit.

[e] A RMSEA value of 0.08 or smaller is considered to indicate good model fit.

[f] A ECV value of 0.6 or larger is considered to indicate sufficient unidimensionality of a model.

Appendices

Appendix 1: Translation process according to FACIT methodology

1. Two independent forward translations (German-English bilingual, native German speakers [MR, SN]);
2. Independent review of resulting translated version of the item bank by a third professional (German-English bilingual, native German speaker), which was followed by a reconciliation meeting involving all translators including an additional researcher who later undertook the cognitive debriefing interviews (SK);
3. Blinded independent back translation of reconciled version (subcontractor, German-English bilingual, native English speaker);
4. Back translation review, i.e., second reconciliation meeting incorporating the back translator's comments;
5. Review and discussion of items with German (n=4), Austrian (n=1), and Swiss (n=1) experts/health professionals to ensure regional appropriateness of the translation;
6. Prefinalization review by project team;
7. Harmonization and quality review by PROMIS translation director (HC);
8. Cognitive debriefing interviews applying the think-aloud technique were used to test understandability and clarity of item wording for patients. For this purpose, the item bank was split into two sets of 60 and 61 items, respectively, as the inclusion of all 121 items in one interview was considered an unreasonably high burden for the participants. Inclusion criteria were age ≥18 years, German language fluency, and a mild to high level of physical impairment. When selecting participants for cognitive debriefing, a balanced sex ratio and the representation of a broad range of age levels were considered to be important. N=10 patients participated in the cognitive debriefing interviews: five with rheumatoid arthritis (50%), three with eating disorders (30%), one with pain disorder (10%), and one with brachial plexus paralysis (10%). The participants' age ranged from 28 to 57 years (mean=39.9, SD=12.5), five (50%) were female. Two participants (20%) were unemployed, one (10%) was retired, and 7 (70%) were employed or self-employed.
9. Modification of problematic items according to cognitive debriefing results;
10. Review and discussion of rewritten items with n=7 patients from Germany and Austria from different clinical settings (GL);

11. Review of rewritten items by above experts/health professionals (see 5.);
12. Finalization of the translated version of the PROMIS PF item bank by consensus, meaning that the PROMIS translation director as well as all members of the international German-speaking group of experts/health professionals (see 5.) had to agree to the final German wording of each individual item;

Appendix 2: Qualitative results of translation and cultural adaptation

Overall, the two forward translations were largely similar. Where differences existed, these could be resolved by discussion. Two major decisions were made by the bilingual expert group (MR, SN, FF, SK) during the reconciliation meeting that have also been mentioned in the results section: First, the English-language item stem 'Are you able to …', prefacing 98 PROMIS v1.2 PF items, was translated with 'Können Sie …', which literally means 'Can you …' but is more straightforward. Second, one item was affected by content-related modifications: Because 'door knobs' are uncommon in German-speaking countries, the German wording for 'door handles' was used for translating this item.

After undergoing the back translation process, further item adjustments and the quality review by the PROMIS translation director (HC), the items were tested as part of cognitive debriefing interviews in n=10 patients: five patients with rheumatoid arthritis, three with eating disorders, one with pain disorder, and one with brachial plexus paralysis were included, with two groups of five each giving feedback to half of the PF items. Participants' age ranged from 28 to 57 years (mean=39.9, SD=12.5), 50% were female. The majority of the translated items were well understood by the patients. The most severe problems appeared for three items, all of which were asking for the ability to climb up one/several/five flight(s) of stairs. The chosen German translation for 'flight of stairs' ('Treppenabsatz') was criticized by all five participants who had trouble understanding the meaning of respective items, i.e. the exact number of steps was unclear to them. To avoid this term, as already described in the results section, we finally used the German wording for 'walking up one/several/five floor(s)' instead. Another particularity in the German translation of the item bank relates to the phrase 'carry a laundry basket'. Resulting from poor item clarity extracted during the cognitive debriefing interviews, this item was translated to 'einen *vollen* Wäschekorb

tragen', which means 'carry a *full* laundry basket'. To ensure that the inherent meaning of respective items was not affected, all item modifications were finalized after consultation with the PROMIS translation director.

A more general issue that arose during the debriefing interviews was that many participants asked for a specific time frame. That is, they argued that they needed a reference such as 'during the past week' or similar to answer the questions adequately. However, the original developers of the PROMIS PF item bank had made the decision to omit a time frame for conceptual reasons. Hence, the question of a potential time frame was not further discussed for the German version.

Appendix 3:

Table A.1: Description and statistics of all items included in any of the PROMIS v1.2 PF short forms

Item code	Content (shortened)	Short forms [a]	Skew-ness	% ceiling / % floor [b]	r_{itc} [c]	H_j [d]
A11	doing chores	SF-4; SF-6; SF-8; SF-10; SF-20	-0.44	32.3/3.4	0.755	0.658
A21	going up and down stairs at a normal pace	SF-4; SF-6; SF-8; Mobility	-0.77	42.1/4.5	0.778	0.659
A23	going for a walk of at least 15 minute	SF-4; SF-6; SF-8; Mobility	-1.82	68.8/2.3	0.776	0.657
A53	running errands and shop	SF-4; SF-6; SF-8	-1.34	59.8/3.4	0.740	0.617
C12	doing two hours of physical labor	SF-6; SF-8; SF-20	0.13	19.2/16.5	0.756	0.713
B1	doing moderate work around the house	SF-6; SF-8	-0.50	36.5/3.8	0.827	0.715
A5	lifting or carrying groceries	SF-8; SF-10; SF-20	-0.27	33.8/3.8	0.830	0.742
A4	doing heavy work around the house	SF-8	-0.04	22.6/12.0	0.804	0.739
A16r1	dressing yourself	SF-10; SF-20; Upper Extremity	-2.08	76.7/0.4	0.746	0.658
A55	washing and drying your body	SF-10; SF-20	-3.20	91.7/0.0	0.724	0.687
B26	shampooing your hair	SF-10; SF-20	-2.79	82.3/0.8	0.709	0.659
C45r1	sitting on and getting up from the toilet	SF-10; SF-20	-2.98	85.0/0.0	0.721	0.689
A1	doing vigorous activities	SF-10; SF-20	0.60	7.5/21.4	0.714	0.734
A3	bending, kneeling, or stooping	SF-10; SF-20	-0.32	33.5/7.5	0.809	0.716
C36r1	walking more than a mile (1.6 km)	SF-10; SF-20	-0.39	34.2/11.3	0.769	0.678
C37	climbing one flight of stairs	SF-10; SF-20	-0.97	52.3/3.4	0.822	0.690
B24	running a short distance	SF-20; Mobility	-0.81	39.1/10.2	0.822	0.691
A38	drying your back with a towel	SF-20; Upper Extremity	-2.68	78.2/1.1	0.674	0.609
B22	holding a plate full of food	SF-20; Upper Extremity	-3.07	83.1/0.8	0.577	0.554
A12	pushing open a heavy door	SF-20	-0.85	47.4/1.5	0.765	0.612
A34	washing your back	SF-20	-1.30	49.8/5.3	0.716	0.594
A51	sitting on the edge of a bed	SF-20	-3.95	90.6/0.0	0.528	0.572
A56	getting in and out of a car	SF-20	-2.29	92.1/0.4	0.772	0.665
B19	squeezing a new tube of toothpaste	SF-20	-4.33	92.9/0.0	0.579	0.642
C46	transferring from a bed to a chair and back	SF-20	-3.09	85.3/0.4	0.699	0.675
A10	standing for one hour	Mobility	-0.24	23.7/15.8	0.682	0.612
A15	standing up from an armless straight chair	Mobility	-1.93	69.9/1.9	0.774	0.662
A31r1	getting up from the floor from lying on back	Mobility	-1.24	56.8/2.3	0.761	0.636
B9	jumping up and down	Mobility	-1.31	57.9/6.8	0.809	0.673
B10	climbing up five steps	Mobility	-2.44	75.6/1.9	0.748	0.659
B32	standing unsupported for 10 minutes	Mobility	-2.01	72.2/2.3	0.745	0.641
B40	standing up on tiptoes	Mobility	-1.61	63.9/4.1	0.710	0.594
B42	standing unsupported for 30 minutes	Mobility	-0.99	51.9/8.3	0.812	0.681
C38	walking at a normal speed	Mobility	-1.67	64.7/4.1	0.805	0.673
B49	going for a short walk (less than 15 minutes)	Mobility	-1.46	65.0/2.6	0.765	0.640
C10	climbing several flights of stairs	Mobility	0.05	27.1/10.2	0.753	0.709
C37	climbing one flight of stairs	Mobility	-0.97	52.3/3.4	0.822	0.690

A17	reaching into a high cupboard	Upper Extremity	-1.70	64.3/1.5	0.799	0.673
A18	using a hammer to pound a nail	Upper Extremity	-2.34	76.6/2.3	0.630	0.558
A20	cutting food using eating utensils	Upper Extremity	-4.32	90.2/0.4	0.491	0.553
A28	opening a can with a hand can opener	Upper Extremity	-2.33	74.4/1.9	0.633	0.554
A29r1	pulling heavy objects towards yourself	Upper Extremity	-1.39	71.8/1.5	0.710	0.595
A35	opening and close a zipper	Upper Extremity	-3.26	88.3/0.0	0.509	0.503
A44	putting on a shirt or blouse	Upper Extremity	-3.32	88.7/0.0	0.618	0.619
A48	peeling fruit	Upper Extremity	-3.60	89.5/0.0	0.551	0.568
A54	buttoning your shirt	Upper Extremity	-3.45	88.7/0.0	0.617	0.622
B21	picking up coins from a table top	Upper Extremity	-3.54	89.1/0.0	0.487	0.484
B30	opening a new milk carton	Upper Extremity	-3.48	85.7/0.8	0.553	0.556
B33	removing something from your back pocket	Upper Extremity	-3.48	88.0/0.0	0.626	0.638
B36	putting on a pullover sweater	Upper Extremity	-2.99	84.6/0.0	0.677	0.650

[a] Item included in one or more PROMIS PF short forms

[b] Percentage who answered item with the highest ('ceiling') or lowest ('floor') possible response category, based on the full sample (n=266)

[c] Corrected item-total correlation: Part-whole corrected correlation between each item and full PROMIS PF scale

[d] Item-specific homogeneity coefficients in the full PROMIS PF item bank