Aus der Medizinischen Klinik mit Schwerpunkt Psychosomatik
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Construct validity of item banking approaches for the assessment of patient-reported outcomes

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Mag. rer. nat. Gregor Liegl
aus Krems, Österreich

Datum der Promotion: 07.12.2018

Content

A. Summary

Abstract (English)

Title

Construct validity of item banking approaches for the assessment of patient-reported outcomes

Background

Item response theory (IRT) methods are increasingly used to standardize the assessment of patient-reported outcomes. By estimating an IRT model with a large number of items measuring the same trait, a construct-based item bank can be established. In theory, any subset of relevant items for a specific population can be selected from an IRT-calibrated item bank to assess an individual's trait level on a standardized scale. However, health-related constructs, such as physical functioning or depression, are often broadly defined, and items of the same item bank may differ in corresponding subdomain or item format, potentially affecting construct validity if different subsets are used for measuring the same latent trait. Based on three studies on recently established item banks, this thesis aims to investigate if different item subsets sufficiently represent the latent construct defined by an entire item bank.

Methods

Study 1: Data from N=3,315 German-speaking subjects who answered the Patient Health Questionnaire depression scale (PHQ-9) were analyzed. For estimating depression scores, PHQ-9 item parameters were used as reported for an IRT-calibrated depression item bank consisting of 143 items from 11 questionnaires in an earlier study. These scores were compared to newly estimated scores resulting from fitting an IRT model solely to the PHQ-9 data. Study 2: The German 121-item PROMIS Physical Function item bank covering different subdomains was psychometrically tested (N=266). Nonparametric IRT and factor analysis were used to evaluate scalability and unidimensionality. Study 3: PROMIS Wave 1 data (N=15,719 subjects from the US) were used to compare measurement precision between three PROMIS Physical Function short forms with similar content but different item format. A common IRT model was estimated for these short forms. Unidimensionality was evaluated using one-factor and bifactor models.

Results

Study 1: Reestimating the model solely based on PHQ-9 data led to similar depression scores compared to using item bank parameters for scoring. Study 2: The PROMIS Physical Function

item bank showed sufficient psychometric properties, including unidimensionality. Scores based on different (subdomain-specific) item subsets were highly correlated with the full item bank. Study 3: The item format affected measurement precision and range but not the underlying construct.

Conclusion

These findings indicate construct validity of using item subsets from large IRT-calibrated item banks for the assessment of patient-reported outcomes. This applies even when the item subsets vary in subdomain-specific content or item format, enabling high flexibility regarding the use of tailored (e.g., population-specific) measurement tools.

Abstrakt (Deutsch)

Titel

Konstruktvalidität von Itembanking-Ansätzen zur Erfassung patientenberichteter Endpunkte

Hintergrund

Methoden der Item-Response Theorie (IRT) werden zunehmend zur standardisierten Erfassung patientenberichteter Endpunkte genutzt. Durch das Schätzen eines IRT-Modells mit einer großen Anzahl an Items, die dieselbe Eigenschaft messen, kann eine konstruktbasierte Itembank kalibriert werden. Theoretisch kann jede Teilmenge einer IRT-kalibrierten Itembank, bestehend aus relevanten Items für eine bestimmte Population („Item-Subset“), verwendet werden, um die Eigenschaftsausprägung einer Person auf einer standardisierten Skala abzubilden. Gesundheitsbezogene Konstrukte, wie körperliche Funktionsfähigkeit oder Depression, sind allerdings oft breit definiert und Items innerhalb einer Itembank können sich hinsichtlich Subdomäne oder des verwendeten Itemformats unterscheiden. Dies könnte die Konstruktvalidität beeinträchtigen, wenn unterschiedliche Item-Subsets zur Erfassung derselben latenten Eigenschaft verwendet werden. Die vorliegende Doktorarbeit umfasst drei Studien zu kürzlich entwickelten Itembanken und hat zum Ziel, zu untersuchen, ob verschiedene Item-Subsets das latente Konstrukt, das durch die Gesamtheit der Items in einer Itembank definiert ist, hinreichend repräsentieren.

Methoden

Studie 1: Daten von N=3,315 deutschsprachigen Personen, die das Depressionsscreening des „Patient Health Questionnaire“ (PHQ-9) beantwortet haben, wurden analysiert. Zur Bestimmung

von Depressionswerten wurden PHQ-9-Itemparameter verwendet, die im Rahmen einer früheren Studie für eine IRT-kalibrierte Depressions-Itembank, bestehend aus 143 Items aus insgesamt 11 Fragebögen, berichtet wurden. Diese Depressionswerte wurden anschließend mit neu geschätzten Depressionswerten verglichen, die aus einem IRT-Modell auf Basis der neuen PHQ-9 Daten resultierten. Studie 2: Die deutschsprachige PROMIS Physical Function Itembank, die verschiedene Subdomänen körperlicher Funktionsfähigkeit umfasst, wurde psychometrisch überprüft (N=266). Nonparametrische IRT-Methoden und Faktorenanalysen wurden verwendet um Skalierbarkeit und Eindimensionalität zu überprüfen. Studie 3: Anhand von PROMIS Wave 1 Daten (N=15,719 Probanden aus den USA) wurde die Messgenauigkeit zwischen drei PROMIS Physical Function Kurzformen mit gleichem Inhalt aber unterschiedlichem Itemformat verglichen. Für die Kurzformen wurde ein gemeinsames IRT-Modell geschätzt. Die Eindimensionalität der Items wurde mittels unidimensionaler Faktorenanalysen und Bifaktor-Modellen überprüft.

Ergebnisse

Studie 1: Die neu geschätzten Depressionswerte auf alleiniger Basis der PHQ-9 Daten waren vergleichbar mit den Depressionswerten auf Grundlage von Itemparametern einer zuvor veröffentlichten Depressions-Itembank. Studie 2: Die deutschsprachige PROMIS Physical Function Itembank zeigte gute psychometrische Eigenschaften, einschließlich Eindimensionalität. Verschiedene (subdomänenspezifische) Item-Subsets korrelierten hoch mit der gesamten Itembank. Studie 3: Das Itemformat beeinflusste Messgenauigkeit und Messbereich, nicht aber das latente Konstrukt.

Fazit

Die Ergebnisse lassen darauf schließen, dass die Erfassung patientenberichteter Endpunkte anhand von Item-Subsets aus umfangreichen IRT-kalibrierten Itembanken konstruktvalide ist. Dies trifft selbst dann zu, wenn sich die Item-Subsets bezüglich der gemessenen Subdomäne oder des Itemformats unterscheiden, was ein hohes Maß an Flexibilität hinsichtlich der Verwendung von maßgeschneiderten (z.B. populationsspezifischen) Messinstrumenten ermöglicht.

1. Introduction

Given the increased burden of chronic and non-communicable diseases in ageing societies, health-related quality of life (HRQoL) has become an important indicator for evaluating the efficacy of healthcare programs and systems [1]. For example, in many high-income countries, the improvement of HRQoL, along with mortality and morbidity, has been defined as the third main criterion for assessing the value of new treatments [2]. Patient-reported outcomes (PRO) for the assessment of HRQoL have consequently become an essential part of medical research [3]. Two of the most frequently assessed HRQoL domains are physical and emotional health [4, 5]. For both, many PRO measures have been developed for different target populations, predominately in the form of fixed-length questionnaires [6, 7]. Most of these traditional measures use instrument-specific scores to assess a person's trait level. Therefore, scores of different measures assessing the same construct cannot be compared on a common scale. This affects the comparability of research results across different diseases and interventions [8]. Moreover, fixed-length instruments bear the limitation that high precision on a wide range of the trait continuum can only be achieved by administering many items, which is burdensome for patients and less practicable in clinical settings [9]. Thus, there is an urgent need for standardized, yet flexible, PRO assessment methods.

The current paradigm shift in PRO measurement away from instrument-based towards construct-based assessment is a promising development with the potential to overcome the above limitations [3]. As a methodological foundation of construct-based assessment, item response theory (IRT) has been used to create construct-specific item banks by calibrating large numbers of items measuring the same latent trait on a common scale [10]. This can be done by estimating a statistical model for the pooled items of existing questionnaires or for a pool of newly written items for a given construct, covering relevant items for all populations of interest. Estimating such an IRT model provides probability functions that describe the relationship between each item and the latent construct. In theory, this has the advantage that the individual response pattern to any item subset can be used to estimate a person's trait level on a standardized metric. This enables the comparison of scores among different legacy measures and tailored (e.g., population-specific) short forms [7, 11], as well as computer-adaptive testing (CAT), i.e., the automatized individualization of item subsets during assessment [12]. Furthermore, once a common IRT metric has been established, measurement precision can be assessed for each item at each trait level, allowing for tailored instruments based on choosing the most informative items for a population of interest [12].

However, although IRT modeling has a long tradition in educational and personality testing [13], PRO item banks still need to prove their worth in clinical practice because of some particularities of HRQoL assessment. First, health constructs are typically broadly defined [14]. For example, items for the assessment of depressive mood can ask about affective, cognitive, behavioral, or somatic aspects of depression; all of which can be included in the same item bank [7]. Second, items of the same item bank can be presented in various item formats (e.g., using different response scales) providing different frames of reference [12]. Therefore, using relatively small item subsets for the assessment of broad constructs as defined by the whole of items included in large item banks may jeopardize construct validity (does the subset assess what it claims to be assessing?).

Using the example of two recently established item banks for depression and physical function, this thesis evaluates the concordance between the latent construct as originally defined by all items of a given full item bank and the specific constructs defined by item subsets that are used for assessing a person's trait level. The value of construct-valid item banks for standardized and flexible PRO assessment is demonstrated. Moreover, the unique opportunity of IRT metrics to compare measurement precision on item level, which is useful for item development and optimization, is illustrated. The findings of this thesis were published in three articles (referred to as paper 1, paper 2, and paper 3; full texts on pp. 25-69).

2. Methods

2.1. Item Bank Development: General Methodological Background

The latent construct and its subdomains need to be clearly specified as the first step in the development of PRO item banks [12]. In accordance with the construct definition, an initial pool of items must be established. This can either be done by pooling the items of existing questionnaires or by creating new items. While the latter approach has some advantages over using existing items (e.g., the possibility to use a consistent response format and to customize the item content to match the construct as well as possible), newly written items need to be pre-tested in debriefing interviews to ensure clarity and comprehensibility for the target population [6]. A useful item bank should include items that cover all subdomains at the full range of trait levels expected in the populations of interest to ensure content-valid and reliable assessment on a wide range of the latent trait continuum [12]. In the example of physical functioning, the item bank should consist of relevant items for all levels of disability for both mobility and upper extremity function.

In the next step, the initial pool of items must be psychometrically evaluated in a sufficiently large and well-distributed (regarding latent trait level) sample to identify items that have the potential to form a common scale and make resulting scores meaningful [9]. It is useful to investigate several psychometric properties prior to fitting a parametric IRT model to eliminate non-fitting items at an early stage. These properties include item skewness, monotonicity, unidimensionality, and local independency; all of which were used for item bank evaluation in this thesis. Highly *skewed items* (e.g., if more than 95% of the sample used the same response category [15]) are of little informational value and should be excluded. *Monotonicity* reflects the assumption that subjects with higher levels on the latent trait are more likely to score higher on an item. Monotonicity can be evaluated by inspecting the item step response functions of a nonparametric IRT model [16]. Items not fulfilling the monotonicity assumptions are not appropriate for calculating meaningful scale scores and should not be retained in the item bank. One-factor confirmatory factor analysis (CFA) is the most widely used method when investigating *unidimensionality*. Due to the categorical nature of most PRO data, CFA models with either a robust weighted least squares means and variance adjusted (WLSMV) estimator or a diagonally weighted least squares (DWLS) estimator have been recommended [17]. Unidimensional model fit can be evaluated by several fit indices. In the studies of this thesis, the comparative fit index (CFI > 0.95 indicating sufficient model fit), the Tucker-Lewis index (TLI > 0.95 indicating sufficient model fit), and the root mean square error of approximation (RMSEA < 0.08 indicating sufficient model fit) were used [7]. If residual correlations between each pair of items did not exceed a value of 0.25, *local independency* was assumed [9]. Local independency is an important assumption of IRT modeling and means that all covariation between the items is explained by the common factor. As traditional one-factor CFA has been discussed as being too restrictive when applied to HRQoL constructs, bifactor models have been recommended as an alternative to evaluate sufficient unidimensionality for IRT modeling [14, 18]. In bifactor models, a general factor represents the common construct of an item bank while several uncorrelated group factors (e.g., one for each subdomain) are allowed. According to Reise et al., a high amount of explained common variance (ECV > 0.6) by the general factor can be used as an indicator of sufficient unidimensionality of a model [19].

For the remaining items with sufficient psychometric properties, an IRT model can be estimated to calibrate all items on a standardized scale. It is important to note that after IRT-modeling further items may need to be excluded, for example due to significant IRT item-fit statistics assessing the discrepancy between observed and model-predicted item responses [20], or due to differential item functioning (DIF) between samples with different diseases or sociodemographic characteristics [12]. The estimation of an IRT model results in logistic functions for each item and describes the

relationship between a person's level on the latent trait and his or her probability of choosing a given response category [12]. The visual representations of these functions are called item characteristic curves (IIC). The IICs can be useful in the inspection of scale and item characteristics [21]. In the past decades, many different IRT models have been developed and used to analyze PRO data [12, 22]. IRT models can vary in several aspects, for example: (i) number of item parameters provided, (ii) number of response categories that can be analyzed, (iii) assumption of ordered response options, (iv) dimensionality assumptions, and (v) how the probabilities of choosing a given response category are calculated [23]. In PRO research, the graded response model (GRM) [24] and the generalized partial credit model (GPCM) [25] are two of the most frequently used IRT models [12]. These models have been applied when establishing and (re-) analyzing the item banks on which this thesis is based [7, 15]. Except for some differences in the model definition, the GRM and the GPCM are very similar and suitable for the same type of data, namely polytomous items (i.e., with more than two response options) with rank ordered response categories measuring a unidimensional construct [12].

In both models, two different kinds of item parameters describing the ICC are estimated: one slope parameter and multiple (the number of response categories minus one) threshold parameters. The *slope* parameter specifies an item's discriminative value and therefore indicates how strong an item is associated with the latent trait (equivalent to factor loadings in CFA). The *threshold* parameters define the difficulty of an item by determining the locations on the latent trait continuum at which an item is most informative, i.e., at which the probability of choosing a given response category is equal to the probability of choosing the adjacent category [21]. Once these parameters are provided, the response pattern to any item subset can be used for estimating latent trait scores [10]. Additionally, the item parameters are useful to identify the item information at each trait level, allowing the most precise items to be chosen in advance if the approximate trait level of the respondent is known [12].

2.2. Item Banks Used for Evaluation

2.2.1. The Common Depression Metric by Wahl et al.

Wahl and colleagues developed a comprehensive item bank by pooling all 143 items of 11 self-report measures for the assessment of depression [7], including the 9-item Patient Health Questionnaire depression scale (PHQ-9) [26]. The included questionnaires differ in many aspects, including number of items, number and wording of response options, and item content. For example, while some questionnaires ask for cognitive and affective indicators of depression only (e.g., loss of interest, depressed mood, concentration difficulties), others also ask for somatic

symptoms, such as sleeping problems or loss of weight. All 143 items were calibrated on a common scale using a GPCM. The model was based on data from several clinical and non-clinical German samples (N=33,844 adults) with scores normed to a T-metric shifted to a general population mean of 50 and a standard deviation (SD) of 10. A total of 54 depression items, including three somatic items of the PHQ-9 (items 3, 5, and 8), did not fulfill the psychometric requirements for fitting a unidimensional IRT model. Therefore, the GPCM was fitted to the 89 remaining items to establish the definitive common depression metric. Consequently, only these 89 items contributed to the final latent depression construct of the item bank. It is noteworthy that none of the well-fitting items asked for somatic depression symptoms. The 54 previously excluded items were subsequently fitted to the common depression metric one by one to provide item parameters for all items of each included depression questionnaire. This was done by estimating separate GPCMs for each non-fitting item together with the 89 well-fitting items, with parameters of the latter being fixed to the values as estimated previously.

Given that many items included in the final item bank did not fit the common depression model (but were subsequently "forced" into the model), including one third of the PHQ-9 items, there is an urgent need to investigate the level of concordance between the latent depression construct of the item bank and the specific depression constructs as defined by the individual questionnaires.

2.2.2. The PROMIS Physical Function Item Bank

The Patient-Reported Outcomes Measurement Information System (PROMIS®) initiative is one of the most extensive projects worldwide integrating rigorous qualitative and quantitative research methods for the development of IRT-based PRO item banks [3]. Funded by the US National Institutes of Health (NIH), PROMIS established many item banks for different HRQoL domains within the past ten years, including a comprehensive physical function (PF) item bank [15]. Aiming to develop a generic PF bank allowing for the precise assessment of physical function on a wide range of ability levels and for use in various clinical and non-clinical populations, PROMIS conducted an extensive item identification and evaluation process, following a standardized data analysis protocol [27]. According to PROMIS, PF is a broad construct defined as an individual's capability "to carry out activities that require physical actions, ranging from self-care (activities of daily living) to more complex activities that require a combination of skills, often within a social context" [28]. A systematic search for PRO measures for the assessment of PF resulted in the identification of 1,860 existing items [6]. After expert evaluation and qualitative analyses, most of these PF items were eliminated as redundant, incomprehensible, unclear, irrelevant, condition-specific, or unrelated to the PROMIS PF construct as defined above. The remaining 168 items

were rewritten to establish a consistent item structure for the item bank. Pre-analyses identified items prefaced with “Are you able to…?” and five response options (“no difficulty” to “unable to do”; Format A) and items prefaced with “Does your health now limit you…?” with five response options (“no difficulty” to “unable to do”; Format B) to reveal adequate psychometric properties and to be the most comprehensible items for participants. Thus, these two item formats were predominately used for item revision. For experimental reasons, some items were presented in a third item format with the item stem “How difficult is it for you to…?” and an extended response scale with six response options (ranging from “very easy” to “impossible”; Format C). The set of 168 revised items was field-tested in healthy and clinical US samples which were part of the PROMIS Wave 1 data collection (N=15,817) [3]. A “block design” was used for collecting data. This means that participants responded to different PF item subsets (“blocks”).

The final PROMIS PF item bank version 1.0 consists of 124 items showing sufficient psychometric properties [15]. Using a GRM, the items were calibrated to a T-metric with an US general population mean of 50 (SD=10). After further optimization, PROMIS PF version 1.2 includes 121 items covering four overlapping subdomains: (1) lower extremity (mobility); (2) central regions (back and neck); (3) upper extremity (grip, reach, and fine-motor control); (4) instrumental activities of daily living (IADL).

2.3. Specific Methods used in Paper 1

The aim of paper 1 (“Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation” [29]) was to evaluate the concordance of the latent depression construct as defined by the IRT-based item bank by Wahl et al. [7] covering the items of 11 depression questionnaires (see 2.2.1) and the specific depression definition of the PHQ-9 [26], which is part of this item bank. Secondary analysis of PHQ data of four German-speaking samples (two from Austria, two from Germany; N=3,315) was conducted. The Austrian samples (n=1006 from general medical practice and orthopedic rehabilitation settings) answered the PHQ-8 depression screening instead of the PHQ-9. The PHQ-8 is equal to the PHQ-9 but omits the ninth item regarding thoughts of suicide. The PHQ-8 has been recommended for screening depressive mood in populations with a low proportion of individuals suffering from major depressive disorder [26].

Unidimensionality and local independence of the PHQ-9 items were assessed using a one-factor CFA model with a WLSMV estimator. Since the PHQ-9 includes items for the assessment of cognitive, affective, and somatic symptoms of depression, a bifactor model was additionally fitted to the data (allowing for three exploratory group factors) to determine the explained common

variance by a general factor representing a common latent depression construct of all items. Sample-related DIF was evaluated using an ordinal logistic regression approach as described by Nagelkerke [30].

PHQ data were analyzed and scored in two different ways. As a first approach, depression scores were estimated for each participant by applying the previously established PHQ-9 item parameters which were reported for the depression item bank by Wahl et al. (based on a common IRT model of 11 questionnaires and data from other samples). As a second approach, a new GPCM was fitted solely to the new data (PHQ items only). To allow for meaningful comparisons across these estimation methods, the newly estimated GPCM was calibrated to the same scale as the common depression metric using two different linking methods. First, mean and covariance of the prior distribution were fixed to the values that resulted from applying the common metric parameters to the data ("model with shifted prior"). Second, Stocking-Lord linking constants were used to linearly transform the newly estimated PHQ-9 item parameters to the common depression metric [31]. Mixed effect models were applied to estimate the effect of estimation method on depression scores. These models included sample and estimation methods as fixed effects and participants as random effects. Pearson correlation coefficients and Bland-Altman plots were used to investigate the agreement in depression scores between the different scoring methods [32].

All analyses were conducted separately for two different sets of items: (1) for all PHQ-9 (or PHQ-8) items and (2) without the three non-fitting PHQ items that were excluded by Wahl et al. for estimating the definitive common depression metric.

2.4. Specific Methods used in Paper 2

In paper 2 ("An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions" [33]), the PROMIS PF item bank was translated into German and psychometrically evaluated in adult patients with a wide range of medical conditions. A stepwise translation and cross-cultural adaptation process was conducted by a bilingual expert group following the Functional Assessment of Chronical Illness (FACIT) methodology [34]. Cognitive debriefing interviews with German-speaking patients were conducted to test the clarity of the items and the conceptual equivalence with the English source. After the translation was finalized, psychometric properties were tested separately for the full PROMIS PF bank (121 items) and different item subsets of this item bank, namely five generic PROMIS PF short forms of different lengths (SF-4a, SF-6b, SF-8b, SF-10a, and SF-20a; ranging from 4 to 20 items) and two subdomain-specific short forms (Mobility, 15 items; Upper Extremity, 16 items) in N=266 patients.

Due to the relatively small sample, parametric IRT was not used to analyze the data. Instead, initial psychometric properties were evaluated using traditional methods as suggested for PROs [35] and nonparametric IRT methods [16]. Sum scores for the PROMIS PF full bank and each short form were calculated according to the PROMIS PF Scoring Manual. To allow for meaningful comparisons of scores across these measures (within-subject), z-score transformation of respective sum scores was applied (standardized to a sample mean of 0; SD=1). Cronbach's alpha and corrected item-total correlations were calculated for each scale to verify internal consistency. To investigate convergent validity, PROMIS PF measures were correlated with the SF-36 physical functioning scale (SF-36 PF-10) [36]. Pearson correlation and the root mean square error (RMSE; indicating the average discrepancy of scores between two measures [37]) were used to investigate the agreement between the full bank and each short form.

Considering the small sample size in relation to the large number of items in the full bank, traditional CFA, using a DWLS estimator, was conducted for the short forms only. Additionally, bifactor models with three exploratory group factors were fitted to each short form. To evaluate unidimensionality of the full PROMIS PF item bank, a nonparametric IRT (NIRT) approach was applied, namely the monotone homogeneity model (MHM) [16]. Loevinger's homogeneity coefficient *H* was used as an indicator of unidimensionality of the data ($H > 0.5$ indicating a strong unidimensional scale) [38].

2.5. Specific Methods used in Paper 3

The objective of paper 3 ("Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function" [39]) was to investigate the effect of different item formats on dimensionality, measurement precision, and measurement range in PROMIS PF items. For this purpose, the data of three experimental 5-item PF short forms, which were included in the PROMIS Wave 1 data collection, were analyzed. These short forms asked about the same activities (doing physical labor, doing yard work, climbing stairs, going for a walk, and opening jars) but used different item formats (Formats A, B, and C as described in 2.2.2).[1]

CFA with a WLSMV estimator was used to evaluate unidimensionality of the experimental items and all remaining items of the final PROMIS v1.0 PF bank, resulting in a total of 134 PF items. Residual correlations between items with similar content (i.e., similar physical activity) but presented in different item format were calculated to investigate local independence. Confirmatory

[1] Only five of the 15 experimental items were included in the final PROMIS v1.0 PF bank (three items using Format A and two using Format B). Format C is not used in the final PROMIS PF bank.

bifactor models with group factors specified for each item format were used as a second approach for evaluating unidimensionality of the items presented in different formats. Due to the block design used for data collection, both the one-factor CFA and the bifactor model had to be fitted to two different subsets of the 134 items.

A GRM was then fitted to the full set of 134 items. Item information functions (IIFs) and test information functions (summarized IIFs of all items in a short form) [12] were calculated to compare measurement precision and measurement range across items and 5-item short forms. These functions were visualized by using item information curves (IICs) and test information curves (TICs), depicting the information provided for each item at each point of the PF continuum. For each IIC and TIC, the area under the curve (AUC) was calculated as an indicator of the overall information provided by a given item or short form.

Because none of the participants rated his or her capability to perform a given activity in all three short forms (due to the block design used for data collection), simulated data were used to directly compare the performance between the format-specific short forms. First, "true" PF scores of five groups with differing average PF levels (n=10,000 respondents per group) were simulated. Next, the corresponding responses to the experimental PF items were simulated based on the item parameters estimated for the GRM. The RMSE between the simulated "true" scores and the scores derived from the short forms was calculated to illustrate format-specific scoring differences. Relative validity (RV) analysis was used to compare the format-specific power to distinguish between groups with different PF levels [40].

3. Results

3.1. Findings of Paper 1

The different CFA fit indices used for evaluating the fit of a one-factor structure of the PHQ-9 and PHQ-8 data led to inconsistent results. While CFIs indicated appropriate model fit in each sample (CFI > 0.95), RMSEA values exceeded the predefined cut-off criterion of 0.08 in three of four samples. However, exploratory bifactor analyses resulted in high explained common variance by the general factor (ECV > 0.74 in each sample), indicating sufficient unidimensionality. The analyses did not indicate a potential problem of local dependency or DIF between subsamples.

The item parameters that resulted from the newly estimated GPCM fitted to the PHQ data were found to be almost identical for the two different linking methods (model with shifted prior and Stocking-Lord linkage). Moreover, fitting the GPCM to the eight PHQ-8 items led to equal

parameters compared to fitting the model to the nine PHQ-9 items. However, the newly estimated PHQ parameters of items 3, 5, and 8 slightly differed from those PHQ parameters that were reported previously for the common depression metric by Wahl et al. [7].

The effect of estimation method on scoring was negligible. Latent depression scores based on the common depression metric parameters and latent depression scores based on fitting a new model solely to the new PHQ data correlated with r > .99 in each sample. When using all PHQ-9 (or PHQ-8) items for scoring, 95% of the individual depression scores differed by less than 1.9 T-scores between estimation methods. When excluding items 3, 5, and 8 for scoring, the agreement across methods was even higher: 95% of the individual depression scores differed by less than 0.8 T-scores between estimation methods.

3.2. Findings of Paper 2

In general, the German version of the PROMIS PF item bank and derived generic and subdomain-specific short forms showed good psychometric properties. Internal consistency was high for the full bank and for each short form (Cronbach's alpha ≥ 0.88). The corrected item-total correlations between each item and the full PROMIS PF bank were found to be sufficiently high and ranged from 0.44 to 0.88 (lowest values were found for hand function items). However, 38% of the sample reached the highest possible score in the 16-item Upper Extremity scale, indicating ceiling effects. Each short form was highly correlated with the full PROMIS PF bank (r=0.87 to 0.99). The correlation of the SF-36 PF-10 with the full PROMIS PF bank (r=0.87) and most short-forms (r=0.80 to 0.90) was very high, except for the Upper Extremity scale (r=0.64). Additionally, Upper Extremity z-scores showed higher discrepancy with full bank z-scores (RMSE=0.54) compared to z-scores derived from other short forms with a similar number of items (e.g., the RMSE was 0.17 for the 20-item SF-20a and 0.28 for the 15-item Mobility scale).

Unidimensionality of each short form was supported by appropriate fit indices found for the one-factor CFA (CFI ranging from 0.996 to 0.998; TLI ranging from 0.995 to 0.998; lower border of the RMSEA 90% CI ranging from 0.036 to 0.071) and by the results of the exploratory bifactor models (ECV ranging from 0.64 to 0.80). Furthermore, sufficient unidimensionality of all 121 items included in the full PROMIS PF bank was indicated by the nonparametric IRT analyses. Loevinger's homogeneity coefficient *H* was above the commonly used cut-off of 0.5 [38] indicating a strong unidimensional scale (*H*=0.646). The monotonicity assumption was not significantly violated for any item.

3.3. Findings of Paper 3

The one-factor CFAs resulted in sufficient fit indices supporting unidimensionality of the 134 PROMIS PF items, including the 15 experimental items in different item formats (CFI and TLI > 0.96; lower border of the RMSEA 90% CI ≤ 0.08). Factor loadings ranged from 0.72 to 0.96 and were very similar for the items that were presented in different formats but asked about the same physical activity. The fit indices of the confirmatory bifactor model were only slightly better (CFI and TLI > 0.97; lower border of the RMSEA 90% confidence interval ≤ 0.074), and loadings on the general factor were substantially higher than the format-specific group factor loadings, indicating sufficient unidimensionality.

IRT analyses did not result in significant item misfit for any experimental short form item. Compared to Formats A and B, items presented in Format C (using an extended 6-category response format) showed the broadest item information curves, indicating high measurement precision on a wider range of the PF continuum. In contrast, items presented in Format B, which was the only item format using a health-related item stem ("Does your health now limit you…?"), led to the highest maximum information on a specific point on the PF continuum. These format-specific differences in item information were consistent across different physical activities. Therefore, test information functions (TIFs) of the format-specific 5-item short forms were affected by the item format in the same way: While the short form using Format B delivered the highest maximum test information, the short form using Format C increased the reliable measurement range (marginal reliability ≥ 0.90) by about half a standard deviation on the positive end of the PF continuum compared to the other short forms. This was also reflected in the total area under the test information curve (AUC), which was considerably larger for Format C.

The simulation study found that Formats A and B allowed for short form T-scores up to 61.8 and 61.0, while the highest possible T-score derived from the short form presented in Format C was 65.5. The agreement between short form scores and simulated "true" scores was similar across all item formats when samples with below-average PF levels were assessed. In contrast, for samples with above-average PF levels, the scoring discrepancy with "true" scores was substantially smaller for Format C short form scores (RMSE ≤ 3.3), compared to using other formats (RMSE up to 4.3 for Format A and 4.4 for Format B). In accordance with these findings, the ability to distinguish between groups with above-average PF levels was significantly better when using the 5-item short form presented in Format C compared to the other formats.

4. Discussion

After applying different approaches to evaluate recently published IRT-calibrated PRO item banks for physical function and depression, this thesis indicates high concordance between the latent constructs defined by the full item banks and the item subsets that were used for scoring.

In the first construct validation approach, the agreement between the item parameters of the PHQ depression scale as reported for a common depression metric and newly estimated item parameters based on PHQ data only was investigated (paper 1). The idea behind this approach was that IRT-based item parameters represent the individual relationship between an item and the latent construct. Consequently, if two different item pools are assessing the same latent construct, this should be reflected by consistent item parameter estimates for those items that are included in both pools, regardless of whether an IRT model is fitted to the one item pool or to the other (provided that both item pools are calibrated on the same scale). A practical consequence is that when using the common items of the two item pools for estimating latent trait scores, there should be no difference whether corresponding item parameters are based on the first or on the second pool of items. Although small differences for some item parameters were found when fitting an IRT model solely to the set of PHQ depression items compared to those parameters that have been reported for a substantially larger depression item bank, the practical impact on estimating depression scores was negligible along the whole range of depression severity. This indicates that both the PHQ depression scale and the common depression metric for 11 instruments have similar underlying construct definitions. Hence, it is possible to simply use item parameters from a common IRT model for estimating depression severity scores that are placed on a common depression metric (and therefore comparable with scores derived from other instruments that are also part of the common metric) without substantially affecting the underlying PHQ depression construct, even though one third of the PHQ-9 items had to be excluded for defining the common depression construct and were fitted to this metric subsequently.

As a second approach to evaluate if all items (and item subsets) of an item bank can be used to assess a common construct, several types of factor analytic methods were used to investigate the unidimensional structure of the PROMIS PF item bank in different samples and languages (papers 2 and 3). In sum, the majority of analyses identified the PROMIS PF items to form a strong unidimensional scale with each item being significantly associated with the latent trait, independent of corresponding subdomain or item format. Therefore, each item subset can be used to assess the underlying physical function construct as defined by PROMIS.

At this point it seems important to point out that according to Bjorner et al. [12], a good PRO measurement instrument should cover all aspects of the latent trait to be assessed to assure content validity. For example, the PROMIS PF item bank has a sufficiently high number of items provided for each of its subdomains [9]. However, the fact that the full item bank is content-valid does not imply that each subset used for assessment is content-valid as well. It has been demonstrated by other authors that items assessing upper extremity are more informative on the lower end of physical function while items of other subdomains are better suited for higher PF levels [15]. This means that developing customized short forms for populations with different PF levels may result in population-specific PF measures with different content. The same applies when computerized adaptive tests (CATs) are used in subjects with different PF levels. CATs usually select those items that provide highest information. Thus, due to the automatized CAT algorithm, patients with low physical function may respond to a high number of upper extremity items, while subjects with above-average physical functioning may not respond to any upper function item at all. Many more assessment scenarios exist that could affect content validity, not least because subdomain-specific short forms are made available by PROMIS ("Pick-a-PRO") and the possibility of selecting the most relevant items for a given (disease-)group ("Build-a-PRO") [15]. However, little is known about the practical consequences of using item subsets that are not content-valid for estimating latent trait scores in the context of comprehensive PRO item banks for the assessment of broad HRQoL constructs. Including at least some items of each subdomain in customized short forms is one way to ensure content validity of item subsets [12]. When administered as CATs, content balancing could be used to ensure that items from each subdomain are selected [41].

As a third approach for evaluating construct validity, the practical consequences of using different item subsets for estimating the latent trait level were investigated by comparing respective scores within subjects (papers 2 and 3). In paper 3, it was found that different item formats used for the assessment of PF, although differing in reference frame (health attribution versus no health attribution) and response scale, did not affect PF scores when respondents with below-average to average PF levels were assessed. For respondents with above-average levels of PF, an item format with an extended response scale (allowing respondents to state that the performance of a physical activity is very easy) resulted in slightly but systematically higher scores; however, this was most likely a result of larger ceiling effects which were apparent when using the other item formats rather than a problem of multidimensionality or low construct validity. It was concluded that the item format did not affect the underlying construct as defined by the PROMIS PF item bank. With respect to item content, the findings of paper 2 indicate that scores derived from generic PF short forms (including items for several PF subdomains) and scores derived from a subdomain-specific

mobility short form showed high agreement with scores derived from the full PROMIS PF item bank. A second subdomain-specific short form for upper extremity functioning showed a somewhat higher scoring discrepancy with the full item bank. However, the correlation between upper extremity and full item bank scores was still high and the lower scoring agreement could have been a result of ceiling effects identified for the upper extremity scale.

In sum, the results indicate that providing common IRT metrics for item banks measuring broad HRQoL constructs offers a flexible, efficient, and construct-valid approach to improve standardized PRO assessment. Thus, this thesis supports the advantages of IRT modeling mentioned in the literature. One advantage is that different item subsets can be used to estimate latent trait scores placed on a standardized scale. As shown by the example of the PHQ-9, this enables researchers and clinicians to directly compare scores across several traditional static questionnaires, which are still the most frequently used instruments for PRO assessment. Moreover, high measurement precision on a wide range of the latent trait continuum can be reached without administering extensive questionnaires with large numbers of items. Instead, customized measurement tools can be used by optimizing item sets for a given study aim, target population, or trait level. This can be done by choosing only the most informative and relevant items for a given population of interest, which is less burdensome for patients and much more feasible in clinical practice. If technical requirements allow, CATs can be used for administering highly individualized item sets, automatically selecting the most informative items during assessment based on previous responses.

A second advantage of establishing a standardized metric based on IRT modelling is that item banks can be extended by adding new items to the common model without changing the metric (i.e., the parameters of the original items). This can be useful when a new fixed-length questionnaire needs be added to a previously established common metric of static PRO measures or when newly developed ceiling or floor items need to be added to an item bank with too few items at the extremes of a trait continuum. Additionally, the possibility of adding items to an existing item bank by fitting a common IRT model allows for the evaluation of the performance of newly developed items. As demonstrated in paper 3, this may be helpful for tuning existing items by modifying characteristics that are not content-related (e.g., item stem wording or number and wording of response options).

5. Limitations

This thesis has several limitations. First, the conclusions are based on only two individual item banks; one for depression and one for physical function. Not only are there more HRQoL constructs that could have been assessed, but there are also more item banks that have been published for both depression and physical function [42, 43]. It cannot be assured that the findings can be generalized to other item banks. Second, the two item banks used for evaluating common IRT metrics differ regarding the construct to be measured as well as the approach used for item bank development. While PROMIS PF items were newly written and non-fitting items were excluded, the common depression metric included pre-existing items and non-fitting items were "forced" into the model (without changing the underlying metric based on the well-fitting items). It is not clear if the findings in one of these item banks are transferable to the other item bank and vice versa. Third, the high concordance between the PHQ-9 depression construct and the latent construct defined by the common depression metric does not imply that this is the case for each of the other questionnaires included in the item bank, which should be evaluated separately in further studies. Finally, the German sample used to evaluate PROMIS PF was rather small and high ceiling effects were indicated for upper extremity items. It is not certain if the lower accordance of z-scores in the upper extremity scale with the full item bank was caused by these ceiling effects or if upper extremity is a different construct than PF as defined by other domains, which has also been discussed by other authors [44]. In this regard, due to the small sample size in relation to the high number of PROMIS PF items, parametric IRT was not used to estimate scores in the German PROMIS PF study. Although the English version was developed by estimating a GRM (indicating that the translated items are likely to fit a common IRT model well), for scoring the German version simple z-score transformations of the scale sum scores had to be used to enable meaningful score comparisons on a common scale. It would be interesting if the scoring discrepancies would still be apparent once IRT item parameters can be used in the German version.

6. Conclusion

In conclusion, this thesis provides evidence that sufficient construct validity can be assumed when using different item subsets of IRT-scaled item banks for the assessment of broad PRO constructs. Based on the findings of three studies, it was illustrated that two general features of IRT modeling also apply to IRT-based assessment of health-related quality of life: (i) the opportunity to directly compare scores across different item subsets (e.g., customized short forms or legacy instruments) and (ii) the possibility to compare measurement precision on item level, which is useful for the development and tuning of items and the optimization of PRO measures.

References

1. Chang S, Gholizadeh L, Salamonson Y, DiGiacomo M, Betihavas V, Davidson PM: **Health span or life span: The role of patient-reported outcomes in informing health policy**. *Health Policy* 2011, **100**(1):96-104.
2. Angelis A, Lange A, Kanavos P: **Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries**. *Eur J Health Econ*, **doi: 10.1007/s10198-017-0871-0**.
3. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S *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**(11):1179-1194.
4. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N: **Patient-Reported Outcomes: The Example of Health-Related Quality of Life—A European Guidance Document for the Improved Integration of Health-Related Quality of Life Assessment in the Drug Regulatory Process**. *Drug Inf J* 2002, **36**(1):209-238.
5. Hand C: **Measuring health-related quality of life in adults with chronic conditions in primary care settings: Critical review of concepts and 3 tools**. *Can Fam Physician* 2016, **62**(7):e375-e383.
6. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, Ware JE, Jr.: **Better assessment of physical function: item improvement is neglected but essential**. *Arthritis Res Ther* 2009, **11**(6):R191.
7. Wahl I, Lowe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, Aita SA, Bergemann N, Brahler E, Rose M: **Standardization of depression measurement: a common metric was developed for 11 self-report depression measures**. *J Clin Epidemiol* 2014, **67**(1):73-86.
8. Puhan MA, Soesilo I, Guyatt GH, Schunemann H: **Combining scores from different patient reported outcome measures in meta-analyses: when is it justified**. *Health Qual Life Outcomes* 2006, **4**:94.
9. Rose M, Bjorner JB, Becker J, Fries JF, 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**(1):17-33.
10. Cella D, Gershon R, Lai J-S, Choi S: **The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment**. *Qual Life Res* 2007, **16**(1):133-141.
11. Fries JF, Cella D, Rose M, Krishnan E, Bruce B: **Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing**. *J Rheumatol* 2009, **36**(9):2061-2066.
12. Bjorner JB, Chang C-H, Thissen D, Reeve BB: **Developing tailored instruments: item banking and computerized adaptive assessment**. *Qual Life Res* 2007, **16**(1):95-108.
13. Embretson SE, Reise SP: **Item response theory**. Mahwah (NJ): Psychology Press; 2000.
14. Reise SP, Morizot J, Hays RD: **The role of the bifactor model in resolving dimensionality issues in health outcomes measures**. *Qual Life Res* 2007, **16 Suppl 1**:19-31.
15. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, 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**(5):516-526.
16. Sijtsma K, Emons WH, Bouwmeester S, Nyklicek I, 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**(2):275-290.
17. DiStefano C, Morgan GB: **A comparison of diagonal weighted least squares robust estimation techniques for ordinal data**. *Struct Equ Modeling* 2014, **21**(3):425-438.
18. Cook KF, Kallen MA, 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**(4):447-460.
19. Reise SP, Scheines R, Widaman KF, Haviland MG: **Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective**. *Educ Psychol Meas* 2013, **73**(1):5-26.
20. Haberman SJ, Sinharay S, Chon KH: **Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions**. *Psychometrika* 2013, **78**(3):417-440.
21. Edelen M, Reeve B: **Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement**. *Qual Life Res* 2007, **16**(1):5-18.
22. Nguyen TH, Han H-R, Kim MT, Chan KS: **An Introduction to Item Response Theory for Patient-Reported Outcome Measurement**. *Patient* 2014, **7**(1):23-35.

23. Chang CH, Reeve BB: **Item response theory and its applications to patient-reported outcomes measurement**. *Eval Health Prof* 2005, **28**(3):264-282.
24. Samejima F: **Graded response model**. In: *Handbook of modern item response theory.* Edited by van der Linden W, Hambleton R. New York: Springer; 1997: 85-100.
25. Muraki E: **A generalized partial credit model**. In: *Handbook of modern item response theory.* Edited by van der Linden W, Hambleton R. New York: Springer; 1997: 153-164.
26. Kroenke K, Spitzer RL: **The PHQ-9: a new depression diagnostic and severity measure**. *Psychiatr Ann* 2002, **32**(9):509-515.
27. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M: **The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years**. *Med Care* 2007, **45**(5 Suppl 1):S3.
28. Hays RD, Spritzer KL, Amtmann D, Lai J-S, DeWitt EM, Rothrock N, DeWalt DA, Riley WT, Fries JF, Krishnan E: **Upper-Extremity and Mobility Subdomains From the Patient-Reported Outcomes Measurement Information System (PROMIS) Adult Physical Functioning Item Bank**. *Arch Phys Med Rehabil* 2013, **94**(11):2291-2296.
29. Liegl G, Wahl I, Berghofer A, Nolte S, Pieh C, Rose M, Fischer F: **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.
30. Nagelkerke NJ: **A note on a general definition of the coefficient of determination**. *Biometrika* 1991, **78**(3):691-692.
31. Stocking ML, Lord FM: **Developing a common metric in item response theory**. *Appl Psychol Meas* 1983, **7**:201-210.
32. Bland JM, Altman DG: **Statistical methods for assessing agreement between two methods of clinical measurement**. *Lancet* 1986, **1**.
33. Liegl G, Rose M, Correia H, Fischer HF, Kanlidere S, Mierke A, Obbarius A, Nolte S: **An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions**. *Clin Rehabil* 2017:269215517714297.
34. Eremenco SL, Cella D, Arnold BJ: **A Comprehensive Method for the Translation and Cross-Cultural Validation of Health Status Questionnaires**. *Eval Health Prof* 2005, **28**(2):212-232.
35. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC: **Quality criteria were proposed for measurement properties of health status questionnaires**. *J Clin Epidemiol* 2007, **60**(1):34-42.
36. Ware JE, Sherbourne CD: **The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection**. *Med Care* 1992, **30**(6):473-483.
37. Stucky BD, Edelen MO, Sherbourne CD, Eberhart NK, Lara M: **Developing an item bank and short forms that assess the impact of asthma on quality of life**. *Respir Med* 2014, **108**(2):252-263.
38. Stochl J, Jones PB, 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):1-16.
39. Liegl G, Gandek B, Fischer HF, Bjorner JB, Ware JE, Rose M, Fries JF, Nolte S: **Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function**. *Arthritis Res Ther* 2017, **19**:66.
40. Deng N, Allison JJ, Fang HJ, Ash AS, Ware JE, Jr.: **Using the bootstrap to establish statistical significance for relative validity comparisons among patient-reported outcome measures**. *Health Qual Life Outcomes* 2013, **11**:89.
41. Zheng Y, Chang C-H, Chang H-H: **Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement**. *Qual Life Res* 2013, **22**(3):491-499.
42. Schalet BD, Revicki DA, Cook KF, Krishnan E, Fries JF, Cella D: **Establishing a Common Metric for Physical Function: Linking the HAQ-DI and SF-36 PF Subscale to PROMIS Physical Function**. *J Gen Intern Med* 2015, **30**(10):1517-1523.
43. Choi SW, Schalet B, Cook KF, Cella D: **Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression**. *Psychol Assess* 2014, **26**(2):513.
44. Hung M, Clegg DO, Greene T, Saltzman CL: **Evaluation of the PROMIS physical function item bank in orthopaedic patients**. *J Orthop Res* 2011, **29**(6):947-953.

B. Eidesstattliche Versicherung

„Ich, Gregor Liegl, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema „*Construct validity of item banking approaches for the assessment of patient-reported outcomes"* selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)" des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.
Meine Anteile an den ausgewählten Publikationen entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.
Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst."

Datum: 23.09.2018 Unterschrift

<u>Anteilserklärung an den erfolgten Publikationen</u>

Gregor Liegl hatte folgenden Anteil an den folgenden Publikationen:

<u>Publikation 1</u>

Gregor Liegl, Inka Wahl, Anne Berghöfer, Sandra Nolte, Christoph Pieh, Matthias Rose, Felix Fischer (2016). Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. *Journal of Clinical Epidemiology* 71:25-34.

Anteil an der Publikation: 75%

- Idee zur multinationalen Studie und Initiierung des Projekts
- Entwicklung des Studiendesigns gemeinsam mit Dr. Felix Fischer
- Zusammenführung und Aufbereitung der zur sekundären Datenanalyse herangezogenen deutschen und österreichischen Datensätze
- Auswertung der Daten und Erstellen von Tabellen und Grafiken in Zusammenarbeit mit Dr. Felix Fischer
- Interpretation der Ergebnisse gemeinsam mit den Koautoren
- Federführung beim Verfassen der Publikation

Publikation 2

Gregor Liegl, Matthias Rose, Helena Correia, H. Felix Fischer, Sibel Kanlidere, Annett Mierke, Alexander Obbarius, Sandra Nolte (2017). 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*, 2017:269215517714297.

Anteil an der Publikation: 80%

- Beteiligung am Übersetzungsprozess der Itembank vom Englischen ins Deutsche
- Durchführung und qualitative Auswertung von Patienteninterviews
- Entwicklung des Studiendesigns zur psychometrischen Überprüfung der Itembank
- Aufbereitung der quantitativen Daten
- Quantitativ-psychometrische Analyse der Daten und Erstellung von Tabellen
- Interpretation der Ergebnisse gemeinsam mit Dr. Sandra Nolte
- Federführung beim Verfassen der Publikation

Publikation 3

Gregor Liegl, Barbara Gandek, Felix Fischer, Jakob B. Bjorner, John E. Ware, Matthias Rose, James F. Fries, Sandra Nolte (2017). Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. *Arthritis Research & Therapy* 19:66.

Anteil an der Publikation: 80%

- Entwicklung des Studiendesigns in Zusammenarbeit mit den Koautoren
- Aufbereitung der Daten und Erstellung des simulierten Datensatzes
- Durchführung der Datenanalyse
- Erstellung von Grafiken und Tabellen
- Interpretation der Ergebnisse in Zusammenarbeit mit den Koautoren
- Federführung beim Verfassen der Publikation

Unterschrift des Doktoranden

_____

C. Printed Versions of Selected Publications

Paper 1: Liegl G, Wahl I, Berghöfer A, Nolte S, Pieh C, Rose M, Fischer F (2016). **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* 71: 25-34.

Impact Factor (2016): 4,978
(Rank 6/90 in HEALTH CARE SCIENCES & SERVICES)

5 Year Impact Factor (2016): 6,939

Eigenfactor (2016): 0,034

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Der Volltext von Paper 1 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1016/j.jclinepi.2015.10.006

Paper 2: Liegl G, Rose M, Correia H, Fischer HF, Kanlidere S, Mierke A, Obbarius A, Nolte S (2017). **An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions**. *Clin Rehabil*: 0269215517714297.

Impact Factor (2016): 2,823
(Rank: 9/65 in REHABILITATION)

5 Year Impact Factor (2016): 3,026

Eigenfactor (2016): 0,007

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Der Volltext von Paper 2 wird in der elektronischen Version meiner Arbeit nicht veröffentlicht. Der Artikel kann unter folgendem Link aufgerufen werden: https://doi.org/10.1177/0269215517714297

Paper 3: Liegl G, Gandek B, Fischer HF, Bjorner JB, Ware JE, Rose M, Fries JF, Nolte S (2017). **Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function**. *Arthritis Res Ther* 19:66.

Impact Factor (2016): 4,121
(Rank 8/30 in RHEUMATOLOGY)

5 Year Impact Factor (2016): 4,537

Eigenfactor (2016): 0,033

Liegl et al. Arthritis Research & Therapy (2017) 19:66
DOI 10.1186/s13075-017-1273-5

Arthritis Research & Therapy

RESEARCH ARTICLE Open Access

Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function

Gregor Liegl[1*], Barbara Gandek[2,3], H. Felix Fischer[1,4], Jakob B. Bjorner[5,6,7], John E. Ware Jr.[2,3], Matthias Rose[1,2], James F. Fries[8] and Sandra Nolte[1,9]

Abstract

Background: Physical function (PF) is a core patient-reported outcome domain in clinical trials in rheumatic diseases. Frequently used PF measures have ceiling effects, leading to large sample size requirements and low sensitivity to change. In most of these instruments, the response category that indicates the highest PF level is the statement that one is able to perform a given physical activity without any limitations or difficulty. This study investigates whether using an item format with an extended response scale, allowing respondents to state that the performance of an activity is easy or very easy, increases the range of precise measurement of self-reported PF.

Methods: Three five-item PF short forms were constructed from the Patient-Reported Outcomes Measurement Information System (PROMIS®) wave 1 data. All forms included the same physical activities but varied in item stem and response scale: format A ("Are you able to …"; "without any difficulty"/"unable to do"); format B ("Does your health now limit you …"; "not at all"/"cannot do"); format C ("How difficult is it for you to …"; "very easy"/"impossible"). Each short-form item was answered by 2217–2835 subjects. We evaluated unidimensionality and estimated a graded response model for the 15 short-form items and remaining 119 items of the PROMIS PF bank to compare item and test information for the short forms along the PF continuum. We then used simulated data for five groups with different PF levels to illustrate differences in scoring precision between the short forms using different item formats.

Results: Sufficient unidimensionality of all short-form items and the original PF item bank was supported. Compared to formats A and B, format C increased the range of reliable measurement by about 0.5 standard deviations on the positive side of the PF continuum of the sample, provided more item information, and was more useful in distinguishing known groups with above-average functioning.

Conclusions: Using an item format with an extended response scale is an efficient option to increase the measurement range of self-reported physical function without changing the content of the measure or affecting the latent construct of the instrument.

Keywords: Physical function, Patient-reported outcomes, Ceiling effects, Measurement range, Item-response theory, Item information, Response scale, Item format

* Correspondence: gregor.liegl@charite.de
[1]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany
Full list of author information is available at the end of the article

Background

Patient-reported outcome (PRO) measures assessing health-related quality of life (HRQoL) have become an essential part of health outcomes research, clinical trials, epidemiological studies, and routine patient monitoring [1–3]. Physical function (PF) is one of the most frequently assessed HRQoL domains [4–6] and has been identified as a core PRO in clinical trials in rheumatic diseases [7]. Thus, efficient assessment of PF is very important. However, traditional PF instruments with a fixed number of items, such as the 10-item Medical Outcome Study Short Form-36 (MOS SF-36®) Health Survey physical functioning scale (PF-10) [8] and the 20-item Health Assessment Questionnaire Disability Index (HAQ-DI) [9], have to compromise between clinical practicality and measurement precision, leading to a limited measurement range on the continuum of physical ability [10].

With the application of item response theory (IRT), any number of items measuring the same latent trait can be calibrated on a common metric. Hence, IRT provides a flexible solution for the challenge of providing practical but still highly precise PRO assessment on a wide range of the latent trait continuum [11–14]. The National Institutes of Health (NIH)-funded Patient-Reported Outcomes Measurement Information System (PROMIS®) has been applying this approach for over 10 years, thereby demonstrating the relevance of IRT item calibration.

PROMIS has developed item banks for a large number of HRQoL domains [2, 15–19], including physical function [10, 20–22]. An important advantage of providing a bank of items scaled on a common metric is that scores derived from different item subsets are directly comparable. This enables the comparison of scores from tailored short forms, which are developed by choosing only the most informative items for a pre-specified trait level and individualized scores from computerized adaptive tests (CATs) [12, 23, 24]. Similarly, if items from different instruments (e.g., short forms) are scaled on the same metric, the measurement precision of these instruments can be directly compared in various populations of interest [25, 26]. This is possible because IRT allows the measurement error of each item (and item subset) to be investigated at each level of the latent trait [27].

Using IRT methods, it has been demonstrated that most PRO instruments measuring PF have satisfactory measurement precision on below average to average functional levels [25, 28]. However, as these instruments have usually been developed for clinical use, they often have ceiling effects in the general population and in samples with higher levels of PF, meaning that a high percentage of these participants achieve the best possible score [29–31]. Thus, individuals with average or above average PF cannot be assessed precisely, leading to low sensitivity to change and larger sample size requirements in clinical trials [28, 29]. The most frequently proposed solution to respond to this shortcoming is the use of items with more difficult content to increase test information on the upper end of a trait continuum [32]. However, this approach might not always be sufficient, e.g., when aiming at extending the measurement range of a static instrument with a fixed number of items or when ceiling effects are still present even after adding new items with more difficult content [33]. In such cases, the modification of the item format of existing items, e.g., by extending the response scale, may present an efficient way of adjusting for ceiling effects [34–36].

Physical function item formats may vary with regard to the item stem, tense (past or present), recall period, attribution (e.g., attribution to health), or response options [4, 35, 37, 38]. For example, in two of the most widely used scales (PF-10, HAQ-DI), the response category that indicates the highest level of PF is the statement that one is able to perform a given activity without any limitations or difficulty [8, 9]. However, there are alternative response scales, for example the one used in the Louisiana State University Health Status Instrument (LSU HSI) [36], that allow respondents to state that the performance of a given activity is easy or even very easy. Such an extended response scale potentially raises the measurement ceiling of PF measures, thus avoiding the necessity of writing new items to measure the ability to perform more difficult activities.

To date, the effect of the item format on item performance in terms of extending the measurement range of PRO measures of PF has not been investigated systematically. To examine the hypothesis that a response format that asks about the ease of doing an activity improves the measurement range, a modification of the LSU HSI item format was incorporated into a set of experimental items in the PROMIS wave 1 data collection [35]. This study uses PROMIS data and IRT to calibrate three five-item short forms with similar content but different item formats on a common metric, to compare the measurement precision and validity of this new item format with two widely used item formats derived from the HAQ-DI and the SF-8™ Health Survey [39].

Methods

Development of the PROMIS PF item bank

To establish the PROMIS PF item bank, a stepwise process integrating qualitative and quantitative item identification and evaluation methods was performed [10, 22, 35], following standard PROMIS procedures [19, 40]. The aim was to develop a generic item bank for use in various patient populations to enable the precise assessment of PF, defined as the capability "to carry out activities that require physical actions, ranging from self-care (activities of daily living) to more complex activities that require a combination of skills, often within a social context" [41].

As detailed elsewhere [35], an initial systematic search for PF instruments resulted in the preliminary retention of 168 unique items, which were rewritten to establish a consistent item structure for the PROMIS item bank. This set of 168 revised items was then field tested in the general population and in clinical samples in the USA (total n = 15,817) and analyzed applying established standard criteria for PROMIS item bank development [39]. To minimize the burden on respondents, items were administered in two different designs: (1) a "full bank" design in which separate subsamples answered either 112 (form C) or 56 (form G) PF items and (2) a balanced incomplete "block" design in which subsamples answered blocks of 21 PF items and items for other PROMIS domains. As a result, each PF item was answered by 2201 to 2926 participants [19, 22]. After psychometric evaluation, the final PROMIS PF item bank version 1.0 consisted of 124 items [22].

Experimental items

Because preparatory analyses showed that the item formats derived from the HAQ-DI [9] (format A: prefaced with "Are you able to …?"; this included five response categories ranging from "without any difficulty" to "unable to do") and the SF-8 [37] (format B: prefaced with "Does your health now limit you …?"; this included five response options ranging from "not at all" to "cannot do") revealed appropriate psychometric properties [10] and appeared to be the formats most comprehensible to participants in a pre-test, these two formats were predominantly used for the aforementioned set of 168 items for field testing [35]. However, for experimental reasons, in a small number of items a modified LSU HSI [36] item format was used (format C: prefaced with "How difficult is it for you …"; this included six response options ranging from "very easy" to "impossible").

To compare the influence of these item formats on item performance, the set of 168 items included 15 experimental items: 5 instrumental activities of daily living (IADLs) of different difficulty levels were presented in all three aforementioned item formats. These three sets of five items differed with regard to the number of response options, definition of the highest and lowest response categories, and attribution to health or not (Table 1). As a result, three five-item short forms with similar content (IADLs) but different item formats were constructed. Of the 15 experimental items, 5 were used in the final 124-item PROMIS PF item bank, with 3 presented in format A and 2 presented in format B.

Data analysis

Item bank evaluation and calibration

Sufficient unidimensionality of the final 124-item PROMIS PF bank had previously been established [22] and was re-evaluated including the 10 additional experimental items, using confirmatory factor analysis (CFA) of a one-factor model with a weighted least squares means and variance

Table 1 Experimental PROMIS PF items for five activities administered in three different item formats

Item format	Item	Item stem	Item content	Number and wording of response options	Attribution to health
A	A1	Are you able to …	… do two hours of physical labor?	5 Without any difficulty 4 With a little difficulty 3 With some difficulty 2 With much difficulty 1 Unable to do	No
	A2[a]		… do yard work like raking leaves, weeding or pushing a lawn mower?		
	A3		… climb several flights of stairs?		
	A4[a]		… go for a walk of at least 15 minutes?		
	A5[a]		… open previously opened jars?		
B	B1[a]	Does your health now limit you in …	… doing two hours of physical labor?	5 Not at all 4 Very little 3 Somewhat 2 Quite a lot 1 Cannot do	Yes
	B2		… doing yard work like raking leaves, weeding or pushing a lawn mower?		
	B3[a]		… climbing several flights of stairs?		
	B4		… going for a walk of at least 15 minutes?		
	B5		… opening previously opened jars?		
C	C1	How difficult is it for you to …	… do two hours of physical labor?	6 Very easy 5 Easy 4 Slightly difficult 3 Difficult 2 Very difficult 1 Impossible	No
	C2		… do yard work like raking leaves, weeding or pushing a lawn mower?		
	C3		… climb several flights of stairs?		
	C4		… go for a walk of at least 15 minutes?		
	C5		… open previously opened jars?		

[a] Item is part of the final Patient Reported Outcomes Measurement Information System Physical Function (PROMIS PF) item bank version 1.0

adjusted (WLSMV) estimator and a bifactor model, specifying local factors for items that shared the same response format. CFA analyses of experimental items in format A used data from "full bank" form C (97 items total), while analysis of formats B and C experimental items used data from "full bank" form G (37 items total); for more information on study design, see [22]. A potential problem of local independence between similar items in Format B and C being administered to the same group was evaluated by analyzing residual correlations. Residual correlation of 0.25 or more was considered potentially problematic and the impact on IRT item parameters was evaluated, as previously described [22].

A graded response model (GRM) was fitted to the set of 134 items consisting of the 15 experimental items (three format-specific short forms) and the remaining 119 items of the final PROMIS PF item bank. Due to the data collection design used for the initial set of 168 PF items, some participants answered only a few of the 134 items analyzed in this study. As in previous analyses [22], only participants who responded to at least two of the 134 PF items were included in the GRM. Although GRM item parameters had already been estimated for the 124 items of the final item bank [22], including 5 of the experimental items, the model was re-estimated to include the 10 additional experimental items. As in previous analyses [22], if a specific response category for an item was answered less than three times, the response option was collapsed with the next higher category to ensure stable item parameter estimates. We estimated item parameters comprising item thresholds and item slopes. Threshold parameters define the range on the latent trait continuum at which a particular response is most likely. The slope parameter specifies the discriminative value of an item. Item fit was evaluated using the $S\text{-}X^2$ statistic.

For estimating individual PF scores, we used the expected-a-posteriori method to calculate theta scores that were subsequently linearly transformed to a *T*-metric (mean = 50, SD = 10 in the calibration sample used in this analysis). To determine the precision of a particular item, we calculated item information functions (IIFs), defining the contribution of an item to the overall precision of the item bank at a given *T*-score level [27]. Differences between IIFs resulting from varying the item format were visualized using item information curves (IICs). Using natural cubic spline interpolation, we calculated the area under the curve (AUC) for each IIC on the empirically observed *T*-score range in the calibration sample as a measure of overall item information. To investigate systematic differences in measurement precision depending on the item format used, we first calculated test information functions for each of the format-specific short forms by summarizing respective IIFs and then we compared the resulting format-specific test information curves and related AUCs.

Simulation study

Due to the study design, no participant in the calibration sample responded to any of the five IADLs used in the experimental items in all three formats. Therefore, to illustrate the performance of all three formats simultaneously, we used simulated data, following the approach used by Voshaar et al. to evaluate PROMIS PF CATs [25]. In the first step, we simulated "true" PF *T*-scores based on the PF score distributions found for five groups in the calibration sample with different self-reported general health; 10,000 "true" PF *T*-scores were simulated for each of the following five general health groups:

(1) Poor general health group:
mean PF *T*-score = 35.6 (SD = 6.5)
(2) Fair general health group:
mean PF *T*-score = 41.9 (SD = 7.6)
(3) Good general health group:
mean PF *T*-score = 48.9 (SD = 7.8)
(4) Very good general health group:
mean PF *T*-score = 54.4 (SD = 7.2)
(5) Excellent general health group:
mean PF *T*-score = 58.8 (SD = 6.5)

In the next step, we simulated responses to the 134 PROMIS PF items for all 50,000 respondents based on their "true" score and the item parameters from the GRM. We scored the three format-specific five-item short forms and the 124-item final PROMIS PF item bank (from now on referred to as the "full bank") using the simulated responses to the respective items in each of these measures.

To illustrate differences in measurement precision due to item format, we calculated root mean square errors (RMSEs) between simulated true scores and corresponding short form scores, with lower values indicating better agreement in estimating individual PF levels [42].

To illustrate how the differences in item format affect the ability to distinguish groups with different levels of PF, we calculated relative validity (RV) coefficients for each format-specific short form [22, 43]. The RV coefficients were calculated using the analysis of variance (ANOVA) *F*-statistic resulting from comparing the full bank PF scores between general health groups as the denominator and the *F*-statistic from comparing short form PF scores between general health groups as the numerator. Hence, the RV coefficient specifies how well a five-item short form with a specific item format distinguishes among groups that differ in PF, compared to using all 124 items of the original PROMIS PF item bank. We calculated 95% confidence intervals for the RV coefficients using standard bootstrap techniques [43, 44]. To provide RV coefficients for different levels of PF, four different general health group comparisons were performed:

(1) Full sample (ANOVA between all five general health groups; n = 50,000)
(2) Average PF compilation (ANOVA between groups with fair, good, and very good general health; n = 30,000)
(3) Below-average PF compilation (ANOVA between groups with poor general health and fair general health; n = 20,000)
(4) Above-average PF compilation (ANOVA between groups with very good and excellent general health; n = 20,000)

CFAs were conducted using Mplus 7.4 [45]. All other statistical analyses were conducted using R 3.1.2 [46]. We used the packages mirt [47] for estimating the GRM and simulating response patterns. For calculating AUCs, we used the package MESS [48]. For plotting item and test information curves, we used ggplot2 [49].

Results

Sample

A total of 15,719 subjects responded to at least two of the 134 items analyzed in this study and therefore were included in the GRM. Of these, only 10 subjects (<0.1%) responded to fewer than 6 items; 99.7% responded to at least 12 items. More than half (54%; n = 8568) responded to one or more of the 15 experimental items (sample characteristics in Additional file 1: Table S1). The experimental items were answered by 2217–2835 participants. The calibration sample had a wide range of PF, with empirically observed *T*-scores (mean = 50, SD = 10) ranging from 11.1 to 73.6.

Evaluation of unidimensionality

Form C and form G had satisfactory fit for the one-factor solution. Factor loadings for the experimental items ranged between 0.83 and 0.93 (format A), 0.83 and 0.96 (format B), and 0.72 and 0.92 (format C). We found residual correlation above 0.25 in one only pair of items (B5 and C5, $r = 0.30$). However, excluding item B5 in the GRM calibration did not notably affect the parameters of item C5 and vice versa, so both items were retained. In the bifactor models, loadings on the global PF factor were substantially higher than loadings on local factors defined by the common response format, thus supporting sufficient unidimensionality of the experimental items and the original PF item bank. For more details, see Additional file 2: Table S2.

Item properties

The results of the IRT analyses for the 15 experimental items (5 IADLs presented in three different item formats) are summarized in Table 2. When adjusting for multiple testing, no item fit-statistic showed significant misfit for any experimental item. Except for one IADL ("open previously opened jars"), item slopes were generally high for all formats. Items prefaced with "Does your health now limit you …" (format B) tended to show slightly higher slope parameters compared to formats A and C (see Table 2).

Item thresholds tended to be similar for format A and format B. In contrast, using format C with the item stem "How difficult is it for you to …" and an extended six-category response scale (ranging from "impossible" to "very easy") expanded the range of the thresholds on the latent trait continuum in both directions. This was particularly pronounced at the positive end of the continuum where the last response in format C increased the measurement range by ≥0.5 SDs of the PF distribution of the sample for all physical activities. As a consequence, the percentage of participants who responded with the highest possible response category was systematically lower (by about 20–25% of the total sample) for items presented in format C compared to the other formats. For two of the more difficult activities (2 hours of physical labor and climbing several flights of stairs), the ceiling effects were halved when using format C compared to both format A and format B (see Table 2).

Figure 1 depicts the IICs for all experimental items presented in different item formats. Format B delivered the highest maximum item information for four of the five physical activities. Moreover, the maximum item information of format B was placed on a systematically higher point on the PF continuum compared to the other formats. In contrast, format C had the broadest measurement range on the *T*-score continuum for each of the five physical activities. The maximum item information of a given item and corresponding points on the latent trait and the AUCs are presented in Table 2. The highest overall item information as specified by the AUC was found for format C except for items asking about opening previously opened jars.

Consequently, the item format affected the total test information provided by the short forms (Fig. 2). The highest maximum test information was found for format B, while items with an extended response format (format C) were highly informative on the widest range on the latent continuum. That is, format C increased the range of highly reliable measurements (defined as marginal reliability ≥0.9 ≈ test information ≥10) by about 0.5 SDs of the PF distribution of the sample on the positive side of the continuum and about 0.1 to 0.2 SDs on the negative side of the continuum.

The cumulative AUC for format C (AUC = 611) was 39% larger than for format A (AUC = 439) and 11% larger than for format B (AUC = 550). When focusing on the item information curve for *T*-scores above 50, the cumulative AUC for Format C (AUC = 192) was 109%

Table 2 Psychometric results for the experimental items presented in three different item formats

Item	Format[b]	Content	Slope	Threshold[c]						Item fit: p (S-X^2)[d]	I_{max} (at T-score)[e]	Area under the curve[f]	Percentage floor/ percentage ceiling[g]
				1	2	3	4	5	Mean				
A1	A	Do 2 hours of physical labor	3.49	38.6	42.9	47.8	54.9		46.1	0.6523	3.71 (T = 42)	92.9	10.4/41.6
B1[a]	B		4.53	38.0	43.0	48.4	53.1		45.6	0.1133	5.93 (T = 49)	132.9	10.0/42.7
C1	C		4.01	37.7	42.0	46.3	52.7	59.8	47.7	0.0358	4.88 (T = 42)	140.3	10.2/19.6
A2[a]	A	Do yard work	4.09	36.3	40.1	44.3	50.7		42.9	0.1473	5.10 (T = 40)	111.1	6.6/57.3
B2	B		4.79	35.7	40.8	46.1	50.6		43.3	0.0751	6.58 (T = 47)	144.0	6.7/52.7
C2	C		4.53	34.3	39.1	43.1	49.3	56.0	44.4	0.0300	6.10 (T = 40)	167.5	5.3/32.1
A3	A	Climb several flights of stairs	3.78	35.2	40.3	45.2	52.0		43.2	0.1722	4.28 (T = 41)	107.0	5.8/51.5
B3[a]	B		4.20	34.2	40.8	46.7	51.3		43.3	0.8460	5.16 (T = 48)	126.0	5.1/51.3
C3	C		3.78	33.3	39.8	44.0	51.0	57.1	45.0	0.1174	4.31 (T = 42)	135.0	6.3/25.4
A4[a]	A	Go for a walk of at least 15 minutes	3.78	33.2	36.4	40.2	45.5		38.8	0.2497	4.45 (T = 37)	91.3	3.7/73.5
B4	B		4.03	32.1	37.2	42.0	45.8		39.3	0.3555	4.93 (T = 43)	107.0	3.4/71.6
C4	C		3.99	30.3	35.6	39.5	44.9	50.8	40.2	0.0033	4.85 (T = 37)	134.7	3.6/47.5
A5[a]	A	Open previously opened jars	1.91	18.8	28.4	37.9			28.4	0.2434	1.10 (T = 28)	36.5	0.9/85.8
B5	B		1.90	12.9	22.8	32.3	39.6		26.9	0.5429	1.10 (T = 33)	39.9	0.3/81.9
C5	C		1.57	5.0	15.5	23.4	34.0	45.4	24.7	0.1877	0.77 (T = 20)	33.6	0.3/62.4

[a]Item is part of the final Patient Reported Outcomes Measurement Information System Physical Function (PROMIS PF) item bank version 1.0. [b]Format A: "Are you able to …" (five-category response scale from "Without any difficulty" to "Unable to do"); format B: "Does your health now limit you in …" (five-category response scale from "Not at all" to "Cannot do"); format C: "How difficult is it for you to …" (six-category response scale from "Very easy" to "Impossible"). [c]Thresholds are transformed to a T-score of 50 ± 10, where 50 = mean and 10 = standard deviation of the analytic sample; slopes are reported unchanged. [d]X^2 statistics (S-X^2) were evaluated after adjusting for multiple testing ($p < 0.0033$). [e]I_{max} (at T-score) depicts the maximum of item information (upper number) of a given item at the corresponding point on the T-score continuum. [f]Total area under the item information curve (IIC) on the empirically observed T-score range in the calibration sample (T-score = 11.1–73.6). [g]Percentage of participants who answered the item with the lowest (floor) or highest (ceiling) possible response category

larger than for format A (AUC = 92) and 81% larger than for format B (AUC = 106).

Agreement between true scores and short forms

The results of the simulation study indicated that the agreement between the simulated true scores and the estimated short form scores was generally lower for formats A and B than for format C (Table 3). Using formats A and B, the agreement with the simulated true scores became even lower when analyzing groups with average to high PF levels (up to RMSE of 4.3 for format A and RMSE of 4.4 for format B). In contrast, the agreement between simulated true scores and short form scores remained relatively constant among all groups when using format C, even in individuals with excellent general health (RMSE ≤3.3).

The highest possible short form T-score was 61.8 when using format A and 61.0 when using format B. In contrast, format C allowed for T-scores up to 65.5, which reduced ceiling effects by more than half in the full simulated sample. Format C was found to be especially beneficial for groups with high PF levels. For example, in the subgroup with "very good" general health, 45.4% of the simulated sample reached the highest possible short form score when using format B. In contrast, only 16.8% of the subgroup with "very good" health reached the highest possible score when using format C. Moreover, lower floor effects were found when using format C.

Distinguishing known groups

The results of the RV analyses using simulated data are presented in Table 4. In most group comparisons (comparisons a, b, and c) the RV was 0.90 or above for all item formats. In contrast, when distinguishing between the two groups with "very good" and "excellent" general health (comparison d), the RV coefficients of format A (RV = 0.79; 95% CI = 0.74–0.84) and format B (RV = 0.78; 95% CI = 0.74–0.83) were considerably lower compared to format C (RV = 0.92; 95% CI = 0.88–0.96).

Discussion

In this study we compared the performance of three different item formats for measuring self-reported PF by analyzing item information. Using simulated data, we illustrated precision in estimating scores and validity in distinguishing between known groups of three five-item short forms with identical content but different item stems and response scales. The five physical activities included in these short forms covered a broad range of item difficulty. Using IRT methodology for data analysis offered the unique opportunity to investigate and visualize measurement precision and range at the item level.

We found strong evidence that the item format may affect the measurement properties of patient-reported PF outcomes. These findings are of practical importance both to researchers and clinicians because this is not

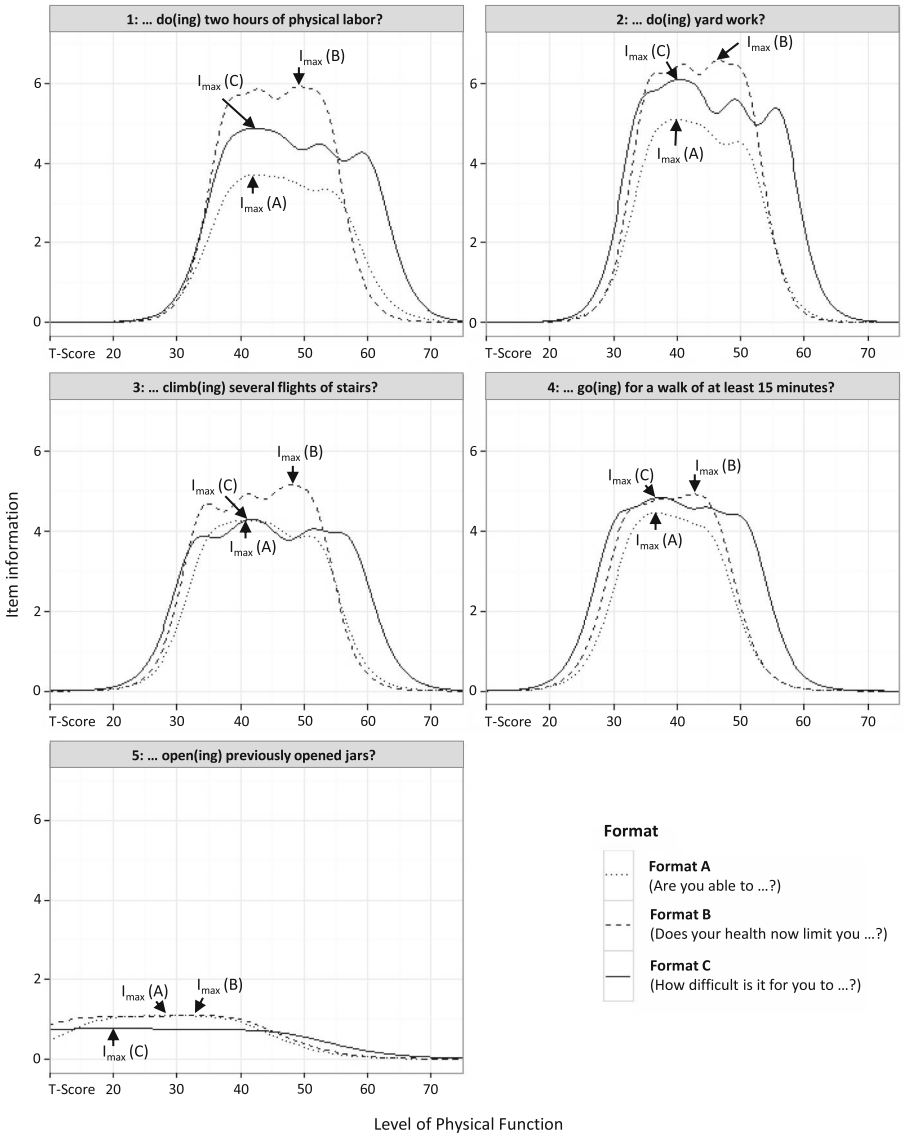

Fig. 1 Comparison of item information functions (IIFs) using different item formats. Format A: "Are you able to …" (five-category response scale from "Without any difficulty" to "Unable to do"); format B: "Does your health now limit you in …" (five-category response scale from "Not at all" to "Cannot do"); format C: "How difficult is it for you to …" (six-category response scale from "Very easy" to "Impossible"). Item parameters and IIFs were initially estimated using a standard normal physical function (PF) metric. PF values were subsequently transformed to a *T*-metric, where 50 = mean and 10 = standard deviations of the analytic sample (*x-axis*). Item information values on the *y-axis* are reported unchanged. I_{max} depicts the specific point on the *T*-score continuum, where a given item delivers maximum item information

only relevant for the development of new instruments but also for the selection of currently available questionnaires for assessing PF in a given population of interest. Moreover, these findings deliver useful information for data interpretation, as the distribution of presumably similar samples can be impacted by the way items are phrased, i.e., identical content but different stem and response format.

In detail, we found that item information differed systematically between the three formats. Format C ("How difficult is it for you to …"), which used an extended response scale including a sixth response option ("very easy"), improved the measurement range by about half a standard deviation on the positive side of the continuum and by about a tenth to a fifth of a standard deviation at the negative end of the continuum, compared to format A ("Are you able to …") and format B ("Does your health now limit you …"). This finding was consistent across different item difficulties. The improvement of the measurement range was found to be particularly beneficial for groups with above-average PF levels, reducing the number of subjects demonstrating ceiling effects in a five-item short

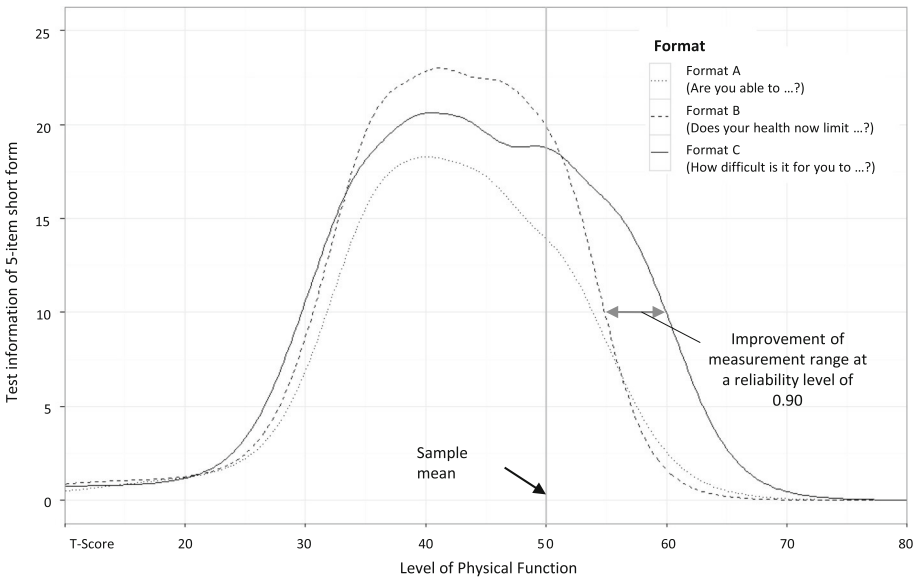

Fig. 2 Comparison of test information functions between different item formats. Format A: "Are you able to …" (five-category response scale from "Without any difficulty" to "Unable to do"); format B: "Does your health now limit you in …" (five-category response scale from "Not at all" to "Cannot do"); format C: "How difficult is it for you to …" (six-category response scale from "Very easy" to "Impossible"). Item parameters and information functions were initially estimated using a standard normal physical function (PF) metric. PF values were subsequently transformed to a *T*-metric, where 50 = mean and 10 = standard deviations of the analytic sample (*x-axis*). Test information values on the *y-axis* are reported unchanged

form by half or even more, when using format C instead of the other item formats. As a consequence, format C was the only item format that had relatively constant measurement precision for all PF levels investigated in the simulation study and had sufficient power to distinguish between groups with above-average functioning. As the improved measurement range of format C was particularly apparent at the positive end of the PF continuum, it seems likely that this improvement was not solely caused by using six instead of five response options but rather by allowing subjects to state that activities were "very easy".

Moreover, our results support that all included item formats measured the same latent construct of PF. The majority of factor loadings were high and their respective magnitude seemed to depend mainly on item content. Consequently, although the final PROMIS PF item bank includes item formats with five-category response options only [35], this study provides evidence that an extended response scale can be applied without affecting the underlying PF construct.

These findings have practical implications for the challenge when encountering ceiling effects, for example, when measuring PF in the general population or in other samples with high PF. The usual way to minimize such ceiling effects is to provide new items with item content that is more relevant for individuals with high PF [32, 33]. However, although providing a larger number of items assessing the extremes of a given trait is undoubtedly useful for the improvement of CATs, this approach does not seem beneficial for increasing the measurement performance of static measures that use the same items for all respondents. Such static measures may still be preferred by many researchers and clinicians for practical reasons [4]. Our findings suggest that it is possible to reduce ceiling effects by optimizing the item format without changing the content of the measures, which may be especially relevant for the future development of items for static PF measures for use in heterogeneous populations with a broad range of ability. However, such modified items should be evaluated psychometrically before use, and additional qualitative item review may be needed. Doing so was beyond the scope of this study.

Another finding of our study is that compared to item formats that do not use attribution, items prefaced with a health-related item stem, as used in format B, delivered the highest maximum item information on a rather narrow range on the PF continuum. Therefore, those types of items seem to be particularly interesting for CATs where highly informative items are selected automatically based on the individual patient's trait level. Moreover, using format B resulted in increased power to distinguish between known groups with close-to-average PF levels compared to the other formats. However, it is not entirely clear if these benefits of format B are caused by health attribution; another reason could be that the wording in format B focuses on "limitations" while both format A and format C ask for "difficulty" in performing physical activities. Further, slightly lower floor effects

Table 3 PROMIS PF full bank and short form scoring characteristics and agreement with simulated "true" scores

General health groups	True PF *T*-score[a] mean (SD)	Full bank (124 items)[b]			Format A[e] (5-item short form)			Format B (5-item short form)			Format C (5-item short form)		
		T-score mean (SD)	RMSE[c]	Percentage floor/ percentage ceiling[d]	*T*-score mean (SD)	RMSE	Percentage floor/ percentage ceiling	*T*-score mean (SD)	RMSE	Percentage floor/ percentage ceiling	*T*-score mean (SD)	RMSE	Percentage floor/ percentage ceiling
Poor	35.6 (6.5)	35.7 (6.4)	0.7	0.0/0.0	36.6 (6.3)	3.0	3.9/0.2	36.4 (6.3)	2.7	1.5/0.2	36.3 (6.3)	2.7	0.5/0.0
Fair	41.9 (7.6)	41.9 (7.7)	0.8	0.0/0.0	42.5 (7.9)	2.9	1.3/3.8	42.3 (790)	2.6	0.5/4.6	42.3 (7.7)	2.5	0.2/0.7
Good	48.9 (7.8)	49.0 (7.9)	1.1	0.0/0.1	49.4 (8.0)	3.2	0.1/17.0	49.4 (8.0)	3.1	0.0/21.5	49.1 (7.9)	2.6	0.0/6.2
Very good	54.4 (7.2)	54.5 (7.3)	1.5	0.0/0.2	54.5 (6.9)	3.8	0.0/37.9	54.5 (6.8)	3.7	0.0/45.4	54.4 (7.3)	2.9	0.0/16.8
Excellent	58.8 (6.5)	58.7 (6.4)	1.9	0.0/0.7	57.8 (5.3)	4.3	0.0/59.0	57.8 (5.0)	4.4	0.0/67.1	58.4 (6.3)	3.3	0.0/32.6
Full sample	47.9 (11.0)	48.0 (11.0)	1.3	0.0/0.2	48.1 (10.4)	3.5	1.1/23.6	48.1 (10.5)	3.4	0.4/27.8	48.1 (10.5)	2.8	0.1/11.3

[a]*T*-scores have a mean of 50 and standard deviation of 10 in the analytic sample. [b]Final Patient Reported Outcomes Measurement Information System Physical Function (PROMIS PF) item bank version 1.0. [c]RSME = root mean square error between estimated *T*-scores and simulated "true" *T*-scores. [d]Percentage of the simulated sample who reached the lowest ("floor") or highest ("ceiling") possible score. [e]Format A: "Are you able to …" (five-category response scale from "Without any difficulty" to "Unable to do"); Format B: "Does your health now limit you in …" (five-category response scale from "Not at all" to "Cannot do"); format C: "How difficult is it for you to…" (six-category response scale from "Very easy" to "Impossible")

Table 4 Analysis of variance (ANOVA) and relative validity (RV)

Subgroup comparisons	General health groups considered for ANOVA[a]					Full bank (124 items)[b]		Format A[c] (5-item short form)		Format B (5-item short form)		Format C (5-item short form)	
	Poor	Fair	Good	Very good	Excellent	F	RV	F	RV[d] (95% CI)	F	RV (95% CI)	F	RV (95% CI)
a. Full sample	X	X	X	X	X	16,957	1 .0	15,582	0.92 (0.91–0.93)	16,139	0.95 (0.94–0.96)	15,712	0.93 (0.92–0.94)
b. Average PF		X	X	X		6960	1 .0	6246	0.90 (0.88–0.91)	6473	0.93 (0.92–0.94)	6349	0.91 (0.90–0.93)
c. Below-average PF	X	X				3818	1 .0	3421	0.90 (0.87–0.92)	3491	0.91 (0.89–0.94)	3564	0.93 (0.91–0.96)
d. Above-average PF				X	X	1870	1 .0	1476	0.79 (0.74–0.84)	1467	0.78 (0.74–0.83)	1720	0.92 (0.88–0.96)

[a]Subgroups marked X were considered for calculating *F* values (ANOVA); n = 10,000 per subgroup. [b]Final Patient Reported Outcomes Measurement Information System Physical Function (PROMIS PF) item bank version 1.0. [c]Format A: "Are you able to …" (five-category response scale from "Without any difficulty" to "Unable to do"); format B: "Does your health now limit you in …" (five-category response scale from "Not at all" to "Cannot do"); format C: "How difficult is it for you to …" (six-category response scale from "Very easy" to "Impossible"). [d]RV calculation: (ANOVA *F* values derived from using a format-specific 5-item short form)/(ANOVA *F* values derived from using full bank scores)

were found for format B (using "cannot do" as the lowest response option) than for format A (using "unable to do" as the lowest response option).

Our study has some limitations. First, our conclusions are based on only five items. Consequently, we cannot be sure that our results apply to all items in the PROMIS PF item bank. However, the format-specific differences were highly consistent among all experimental items. A second limitation concerns the selection of only three item formats. Among PRO instruments for the assessment of PF there is a large variety of item formats, which differ in many more aspects than the response scale and item stem [35, 37, 38]. Future studies should clarify whether other formats should be considered for further optimization of measurement precision, and also if the wording of the formats used in this study can be further improved [50]. In particular, modifications might be made to format C, which is based on the LSU HSI (format C: "How difficult is it for you to …"), in which the item stem asked about difficulty but not ease, whereas the corresponding response set included "easy" and "very easy".

Third, we had to use simulated data for illustrating differences in measurement precision due to the item formats because the study design did not permit direct comparisons using real data. Fourth, it has been shown that PF measures are not only limited by ceiling effects but also by floor effects when assessing highly disabled populations [33]. It seems unlikely that this issue can be solved sufficiently by simply modifying the response scale, as the most extreme response option at the negative end of the trait continuum is usually rated "impossible". For highly disabled samples, it may therefore be necessary to include items asking about basic activities of daily living (ADLs). Finally, although we found differences in measurement precision between the item formats, it remains unclear whether one of the formats used in this study is superior to the others in measuring what a person is actually able to perform, i.e., as measured by performance-based outcome measures.

Conclusions

This study systematically investigated differences in measurement properties resulting from extending the response scale of PRO measures assessing PF. Our findings provide evidence that using an extended six-category response format, including the response options "easy" and "very easy", is an efficient and valid way to considerably extend the range of precise measurement of PF at the positive end of the trait continuum without changing the content of the measure or affecting the latent construct of the instrument. Optimizing the item format offers an effective opportunity to improve measurement precision and to reduce ceiling effects. This is especially relevant for the application of generic short forms in populations with average and above-average levels of PF and for the selection of global items measuring PF.

Additional files

Additional file 1: Table S1. Summary of sample characteristics. (DOCX 38 kb)

Additional file 2: Table S2. Results of confirmatory factor analyses. (DOCX 38 kb)

Abbreviations

ADL: Basic activities of daily living; ANOVA: Analysis of variance; AUC: Area under the curve; CAT: Computerized adaptive testing; CFA: Confirmatory factor analysis; GRM: Graded response model; HAQ-DI: Health Assessment Questionnaire Disability Index; HRQoL: Health-related quality of life; IADL: Instrumental activities of daily living; IIC: Item information curve; IIF: Item information function; IRT: Item response theory; LSU HSI: Louisiana State University Health Status Instrument; MOS SF-36: Medical Outcome Study Short Form-36; PF: Physical function; PF-10: Medical Outcome Study Short Form-36 Health Survey Physical Function scale; PRO: Patient-reported outcome; PROMIS: Patient-reported outcomes measurement information system; RMSE: Root mean square error; RV: Relative validity

Acknowledgements

Not applicable.

Funding
Data analysis and preparation of the article was supported by a Rahel-Hirsch scholarship from the Charité - Universitätsmedizin Berlin to SN. Additional support was provided by University of Massachusetts Medical School from its own research funds. This article uses data collected and developed under the Patient Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org), which was funded by the National Institutes of Health (NIH) Common Fund Initiative under a number of cooperative agreements, including an agreement with Stanford University (PI: James Fries, MD, U01AR52158) to develop the Wave 1 PF item bank and with Northwestern University (PI: David Cella, PhD, U01AR52177) for the PROMIS Statistical Co-ordinating Center. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders did not have any role in study design, in the analysis and interpretation of data, in the writing of the manuscript, or in the decision to submit the article for publication.

Availability of data and materials
The PROMIS wave 1 dataset analyzed during the current study is available in the Harvard Dataverse repository (https://dataverse.harvard.edu). The set of simulated data generated and analyzed during the current study is available from the corresponding author (GL) on reasonable request.

Authors' contributions
GL contributed to study conception and design, analyzed and interpreted the data, wrote the first draft of the manuscript, and had primary responsibility for manuscript revision. BG contributed to study conception and design, data analysis and interpretation, and revised the manuscript. FF conducted statistical analyses, and contributed to interpretation of the data and manuscript revision. JBB contributed to study conception and design, data analysis and interpretation, and manuscript revision. JEW and MR contributed to study conception and design, interpretation of the data, and manuscript revision. JFF contributed to study conception and design and manuscript review. SN contributed to study conception, statistical analyses, interpretation of the data, and manuscript conception and revision.
All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The data collection was approved by the Institutional Review Boards at Northwestern University (for the Statistical Coordinating Center) and Stanford University. All participants provided written informed consent.

Author details
[1]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. [2]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA. [3]John Ware Research Group, Watertown, MA, USA. [4]Institute for Social Medicine, Epidemiology and Health Economics, Charité – Universitätsmedizin Berlin, Berlin, Germany. [5]National Research Centre for the Working Environment, Copenhagen, Denmark. [6]Optum, Lincoln, RI, USA. [7]Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [8]Department of Immunology and Rheumatology, Stanford University School of Medicine, Palo Alto, CA, USA. [9]Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Melbourne, VIC, Australia.

Received: 10 October 2016 Accepted: 27 February 2017
Published online: 21 March 2017

References

1. Ahmed S, Berzon RA, Revicki DA, Lenderking WR, Moinpour CM, Basch E, et al. The use of patient-reported outcomes (PRO) within comparative effectiveness research: implications for clinical practice and health care policy. Med Care. 2012;50:1060–70.
2. Garcia SF, Cella D, Clauser SB, Flynn KE, Lad T, Lai J-S, et al. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. J Clin Oncol. 2007;25:5106–12.
3. Calvert M, Thwaites R, Kyte D, Devlin N. Putting patient-reported outcomes on the 'Big Data Road Map'. J R Soc Med. 2015;108:299–303.
4. Schalet BD, Revicki DA, Cook KF, Krishnan E, Fries JF, Cella D. Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS physical function. J Gen Intern Med. 2015;30:1517–23.
5. Kluetz PG, Slagle A, Papadopoulos EJ, Johnson LL, Donoghue M, Kwitkowski VE, 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.
6. Oude Voshaar MA, ten Klooster PM, Taal E, Krishnan E, van de Laar MA. 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;14:1–7.
7. van Tuyl LH, Boers M. Patient-reported outcomes in core domain sets for rheumatic diseases. Nat Rev Rheumatol. 2015;11:705–12.
8. Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30(6):473–83.
9. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol. 2003;30:167–78.
10. Rose M, Bjorner JB, Becker J, Fries JF, 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.
11. Embretson SE, Reise SP. Item response theory. Mahwah (NJ): Psychology Press; 2000.
12. Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res. 2007;16:133–41.
13. Liegl G, Wahl I, Berghofer A, Nolte S, Pieh C, Rose M, 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.
14. Petersen MA, Aaronson NK, Arraras JI, Chie WC, Conroy T, Costantini A, et al. The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency. J Clin Epidemiol. 2013;66:330–9.
15. Amtmann D, Cook KF, Jensen MP, Chen W-H, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010;150:173–82.
16. Hahn EA, DeVellis RF, Bode RK, Garcia SF, Castel LD, Eisen SV, et al. Measuring social health in the patient-reported outcomes measurement information system (PROMIS): item bank development and testing. Qual Life Res. 2010;19:1035–44.
17. Lai J-S, Cella D, Choi S, Junghaenel DU, Christodoulou C, Gershon R, 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;92:20–7.
18. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment. 2011;18:263–83.
19. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, 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.
20. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and PROMIS: assessment of physical function. J Rheumatol. 2014;41:153–8.
21. Oude Voshaar MA, Ten Klooster PM, Glas CA, Vonkeman HE, Taal E, Krishnan E, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity-driven 20-item short form in rheumatoid arthritis compared with traditional measures. Rheumatology (Oxford). 2015;54:2221–9.
22. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware Jr JE. 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.

23. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol. 2005;23:53–7.
24. Ware Jr JE, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlöf CG, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res. 2003;12:935–52.
25. Oude Voshaar MAH, ten Klooster PM, Glas CAW, Vonkeman HE, Krishnan E, van de Laar MAFJ. 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.
26. Wahl I, Lowe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014;67:73–86.
27. Bjorner JB, Chang C-H, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res. 2007;16:95–108.
28. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011;13:R147.
29. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol. 2011;38:1759–64.
30. Oude Voshaar MA, ten Klooster PM, Taal E, van de Laar MA. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes. 2011;9:99.
31. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis. 1995;54:461–5.
32. Bruce B, Fries J, Lingala B, Hussain YN, Krishnan E. Development and assessment of floor and ceiling items for the PROMIS physical function item bank. Arthritis Res Ther. 2013;15:R144.
33. Fries JF, Lingala B, Siemons L, Glas CA, Cella D, Hussain YN, et al. Extending the floor and the ceiling for assessment of physical function. Arthritis Rheumatol (Hoboken, NJ). 2014;66:1378–87.
34. Marfeo EE, Ni P, Chan L, Rasch EK, Jette AM. Combining agreement and frequency rating scales to optimize psychometrics in measuring behavioral health functioning. J Clin Epidemiol. 2014;67:781–4.
35. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther. 2009;11:R191.
36. Fisher Jr WP, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI physical functioning scales. J Outcome Meas. 1997;1:329–62.
37. Holsbeeke L, Ketelaar M, Schoemaker MM, Gorter JW. Capacity, capability, and performance: different constructs or three of a kind? Arch Phys Med Rehabil. 2009;90:849–55.
38. Young NL, Williams JI, Yoshida KK, Bombardier C, Wright JG. The context of measuring disability: does it matter whether capability or performance is measured? J Clin Epidemiol. 1996;49:1097–101.
39. Ware J, Kosinski M, Dewey J, Gandek B. How to score and interpret single-item health status measures: a manual for users of the SF-8 health survey. Lincoln: QualityMetric Incorporated; 2001.
40. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45:S22–31.
41. PROMIS: Dynamic tools to measure health outcomes from the patient perspective. Available at: http://www.nihpromis.com/Measures/domainframework1. Accessed 7 Mar 2017.
42. Stucky BD, Edelen MO, Sherbourne CD, Eberhart NK, Lara M. Developing an item bank and short forms that assess the impact of asthma on quality of life. Respir Med. 2014;108:252–63.
43. Deng N, Allison JJ, Fang HJ, Ash AS, Ware JE. Using the bootstrap to establish statistical significance for relative validity comparisons among patient-reported outcome measures. Health Qual Life Outcomes. 2013;11:89.
44. Efron B. Better bootstrap confidence intervals. J Am Stat Assoc. 1987;82:171–85.
45. Muthén LK, Muthén BO. Mplus User's Guide. CA: Muthén & Muthén; 1998-2015
46. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0; 2014.
47. Chalmers RP. mirt: A multidimensional item response theory package for the R environment. J Stat Softw. 2012;48:1–29.
48. Ekstrom C, Ekstrom MC. Package 'MESS'. 2012.
49. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer Science & Business Media; 2009.
50. Dillman DA, Smyth JD, Christian LM. Internet, phone, mail, and mixed-mode surveys: the tailored design method. Hoboken (NJ): Wiley; 2014.

Additional file 1: Table S1. Summary of sample characteristics

	Full sample [a]	Experimental item subsample [b]
n	15,719	8,568
Mean age (SD)	53.85 (16.6)	54.66 (16.8)
% Female	52.1	49.5
% White	82.7	82.5
% > High School graduate	82.1	82.0
% Any chronic condition [c]	77.6	76.4
% Musculoskeletal [d]	35.5	29.7
% Cardiopulmonary [d]	33.5	34.0
% Mental disorder [d]	33.5	26.0
% Cancer [d]	19.7	27.1
% Gastroenterological [d]	17.7	17.6
% Neurological [d]	5.9	2.3

[a] All participants who responded to at least two of the 134 items analyzed in the present study

[b] Subsample of participants who responded to at least one of the experimental items presented in different formats

[c] Participants were asked for current or previous chronic conditions: 'Have you ever been told by a doctor or a health professional that you have …'

[d] Participants who reported musculoskeletal disorders (osteoarthritis (OA), rheumatoid arthritis (RA)), cardiopulmonary conditions (asthma, chronic obstructive pulmonary disease (COPD), coronary heart disease, myocardial infraction, heart failure), mental disorders (depression, anxiety, addiction), cancer of any origin, gastroenterological disorders (diabetes, liver disease), or neurological disorders (amyotrophic lateral sclerosis, multiple sclerosis, spinal cord injury (SCI), Parkinson's disease). Percentages sum to over 100% as most participants reported more than one condition.

Additional file 2: Table S2. Results of Confirmatory Factor Analyses

	One-factor model		Bifactor model	
	Form C [a] *n = 639*	Form G [b] *n = 757*	Form C [a] *n = 639*	Form G [b] *n = 757*
Chi-square	7491 (df = 4559), p < 0.001	3869 (df = 629), p < 0.001	6182 (df = 4462), p < 0.001	3017 (df = 592), p < 0.001
CFI	0.982	0.970	0.989	0.977
TLI	0.982	0.968	0.989	0.975
RMSEA	0.032 [0.030 - 0.33]	0.083 [0.080 - 0.085]	0.025 [0.023 - 0.026]	0.074 [0.071 - 0.076]

RMSEA: Root Mean Square Error of Approximation; CFI: Comparative Fit Index; TLI: Tucker Lewis Index

[a] Set of 97 items consisting of 5 experimental items using Format A: 'Are you able to…' (Five-category response scale from 'Without any difficulty' to 'Unable to do') and 92 other items of the final PROMIS PF item bank; n=639 subjects responded to all of these items

[b] Set of 37 items consisting of 10 experimental items using Format B: 'Does your health now limit you in…' (Five-category response scale from 'Not at all' to 'Cannot do') and Format C: 'How difficult is it for you to…' (Six-category response scale from 'Very easy' to 'Impossible') and 27 other items of the final PROMIS PF item bank; n=757 subjects responded to all of these items.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

E. Complete List of Publications

Peer-reviewed Publications

Obbarius N, Fischer F, Obbarius A, Nolte S, **Liegl G**, Rose M (2018). A 67-item stress resilience item bank showing high content validity was developed in a psychosomatic sample. *J Clin Epidemiol* 100:1-12.

Obbarius A, Obbarius N, Fischer F, **Liegl G**, Rose M (2018). [Evaluation of Factor Structure and Construct Validity of the 12-Item Short Version of the OPD Structure Questionnaire in Psychosomatic Patients]. *Psychother Psych Med*, doi: 10.1055/s-0043-125394

Jank R, **Liegl G**, Böckle M, Vockner B, Pieh C (2017). [Frequency of somatic syndromes in primary care]. *Z Psychosom Med Psychother* 63(2):202-12.

Liegl G, Rose M, Correia H, Fischer HF, Kanlidere S, Mierke A, Obbarius A, Nolte S (2017). An initial psychometric evaluation of the German PROMIS® v1.2 Physical Function item bank in patients with a wide range of health conditions. *Clin Rehabil*: 0269215517714297.

Liegl G, Gandek B, Fischer HF, Bjorner JB, Ware JE, Rose M, Fries JF, Nolte S (2017). Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. *Arthritis Res Ther* 19:66.

Klokker L, Terwee CB, Wæhrens EE, Henriksen M, Nolte S, **Liegl G**, Kloppenburg M, Westhoven R, Wittoek R, Kjeken I, Haugen IK, Schalet B, Gershon R, Bliddal H, Christensen R (2016). Hand-related physical function in rheumatic hand conditions: a protocol for developing a patient-reported outcome measurement instrument. *BMJ open* 6(12):e011174.

Fischer HF, Wahl I, Nolte S, **Liegl G**, Brähler E, Löwe B, Rose M (2016). Language-related differential item functioning between English and German PROMIS Depression items is negligible. *Int J Methods Psychiatr Res*, doi: 10.1002/mpr.1530.

Boeckle M, **Liegl G**, Jank R, Pieh C (2016). Neural correlates of conversion disorder: overview and meta-analysis of neuroimaging studies on motor conversion disorder. *BMC psychiatry* 16(1):195.

Liegl G, Boeckle M, Leitner A, Pieh C (2016). A meta-analytic review of brief guided self-help education for chronic pain. *Eur J Pain* 20(10):1551-62.

Fischer KI, **Liegl G**, Rose M, Nolte S (2016). [The measurement of health-related quality of life using modern test theory methods - Development and application of computer adaptive tests]. *Pflege & Gesellschaft* 21(2):130-44.

Liegl G, Wahl I, Berghöfer A, Nolte S, Pieh C, Rose M, Fischer F (2016). 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* 71:25-34.

Boeckle M, Schrimpf M, **Liegl G**, Pieh C (2016). Neural correlates of somatoform disorders from a meta-analytic perspective on neuroimaging studies. *NeuroImage: Clinical* 11:606-13.

Boeckle M, Chetouani Y, Schrimpf M, **Liegl G**, Leitner A, Pieh C (2015). [Austrian expenditures on pharmaceutical drugs between 2006 and 2013]. *Z Psychosom Med Psychother* 61(4):359-69.

Schrimpf M, **Liegl G,** Boeckle M, Leitner A, Geisler P, Pieh C (2015). The effect of sleep deprivation on pain perception in healthy subjects: a meta-analysis. *Sleep Med* 16(11):1313-20.

Liegl G, Plessen CY, Boeckle M, Leitner A, Pieh C (2015). Guided self-help interventions for irritable bowel syndrome: a systematic review and meta-analysis. *Eur J Gastroenterol Hepatol* 27(10):1209-21.

Boeckle M, **Liegl G**, Leitner A, Pieh C (2014). How burdensome is the treatment of patients with somatoform disorders? *Z Psychosom Med Psychother* 60(4), 383-91.

Leitner A, Märtens M, Koschier A, Gerlich K, **Liegl G**, Hinterwallner H, Schnyder U (2013). Patients' perceptions of risky developments during psychotherapy. *J Contemp Psychother*, 43(2), 95-105.

Märtens M, **Liegl G** (2013). [Patient rights act in a psychotherapeutic context. Requirements for informed consent and treatment alternatives]. *Psychotherapeut* 58:73-8.

Book Chapters

Märtens M, Koschier A, **Liegl G** (2014). Beziehung gut, Ende gut - stimmt nicht immer: Die Qualität der therapeutischen Beziehung und ihr Einfluss auf Behandlungszufriedenheit und Dauer der Therapie. In A. Leitner, B. Schigl, M. Märtens (Ed.), *Wirkung, Risiken und Nebenwirkungen von Psychotherapie* (pp. 38-45). Wien: Facultas. ISBN: 3708911253

Liegl G, Leitner A (2014). Grenzmethoden in der Psychotherapie: Konstruktives Eingehen auf menschliche Bedürfnisse oder Nährboden für Risiken und Nebenwirkungen?. In A. Leitner, B. Schigl, M. Märtens (Ed.), *Wirkung, Risiken und Nebenwirkungen von Psychotherapie* (pp. 66-69). Wien: Facultas. ISBN: 3708911253

Leitner A, **Liegl G**, Märtens M (2014). Unterschiedliche Verfahren - unterschiedliche Risiken. In A. Leitner, B. Schigl, M. Märtens (Ed.), Wirkung, Risiken und Nebenwirkungen von Psychotherapie (pp. 46-50). Wien: Facultas. ISBN: 3708911253

Liegl G, Koschier A (2014). Selbsterfahrung - brauche ich das? Der subjektiv wahrgenommene Bedarf an Selbsterfahrung in der Psychotherapieausbildung in Österreich. Wann sind wir gut genug? In S. B. Gahleitner, R. Reichel, B. Schigl, A. Leitner (Ed.), *Selbstreflexion, Selbsterfahrung und Selbstvorsorge in Psychotherapie, Beratung und Supervision* (pp. 91-105). Weinheim und Basel: Beltz Juventa. ISBN: 978-3-7799-2925-3

Conference Papers Published in Peer-reviewed Journals (as first author)

Liegl G, Kanlidere S, Stengel A, Obbarius A, Knebel F, Buttgereit F, Rose M, Nolte S (2016). Generic self-reported and performance-based measures of physical function are highly correlated but differentially affected by pain and illness perception. *Qual Life Res* 25(1 Supplement):1-196.
Conference: 22nd Annual Conference of the International Society for Quality of Life Research, Copenhagen, Denmark; 10/2016

Liegl G, Kanlidere S, Knebel F, Obbarius A, Stengel A, Buttgereit F, Rose M, Nolte S (2016). Prior execution of physical tasks influences patients' self-assessment of physical function. *Qual Life Res* 25(1 Supplement):1-196.
Conference: 22nd Annual Conference of the International Society for Quality of Life Research, Copenhagen, Denmark; 10/2016

Liegl G, Fischer F, Bjorner JB, Fries JF, Gandek B, Ware J, Nolte S, Rose M (2015). The impact of item format on item information and resulting person parameters in patient-reported outcomes measuring physical function. *Qual Life Res* 24(1 Supplement):1-191.

Conference: 22nd Annual Conference of the International Society for Quality of Life Research, Vancouver, British Columbia, Canada; 10/2015

Liegl G, Wahl I, Berghöfer A, Nolte S, Pieh C, Rose M, Fischer F (2015). Applying a common depression metric to independent samples: Different approaches result in similar score estimates. *Qual Life Res* 24(1 Supplement):1-191.
Conference: 22nd Annual Conference of the International Society for Quality of Life Research, Vancouver, British Columbia, Canada; 10/2015

Liegl G, Boeckle M, Pieh C (2014). Prevalence of common mental disorders in primary care in Austria. *J Psychosom Res* 76(6):509.
Conference: Annual Meeting of the European Association of Psychosomatic Medicine, Sibiu, Romania; 06/2014

Liegl G, Boeckle M, Pieh C (2014). Influence of depression, expectation of therapy effectiveness, and self-efficacy on the treatment outcome in patients with multiple somatic symptoms (MSS). *J Psychosom Res* 76(6):508.
Conference: Annual Meeting of the European Association of Psychosomatic Medicine, Sibiu, Romania; 06/2014

F. Acknowledgements

An dieser Stelle möchte ich all jenen Menschen ausdrücklich danken, ohne die diese Doktorarbeit nicht oder nur wesentlich erschwert möglich gewesen wäre.

Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. Matthias Rose sowie meiner Promotionsbetreuerin Dr. Sandra Nolte für die fachliche und motivationale Unterstützung, ihre kontinuierliche konstruktive Kritik, ihr Vertrauen in meine wissenschaftlichen Kompetenzen und für die gewährte Arbeitsautonomie bei der Erstellung der Publikationen, die dieser Dissertation zugrunde liegen.

Des Weiteren möchte ich mich bei Dr. Felix Fischer für den häufigen fachlichen Austausch und seine großartige Unterstützung in jeglichen statistischen Belangen bedanken.

Bei dem gesamten Team unserer Health Outcomes Research Arbeitsgruppe an der Charité – Universitätsmedizin Berlin bedanke ich mich für viele konstruktive Diskussionen und die angenehme und freundschaftliche Arbeitsatmosphäre. Alex, Kathrin, Nina, Paul, Iris, Andrea: Dank euch waren die letzten drei Jahre nicht nur durch konstruktive Zusammenarbeit gekennzeichnet, sondern auch durch Spaß und viel Freude am gemeinsamen Tun.

Danken möchte ich außerdem meinen Eltern Monika Höllebauer und Walter Liegl sowie meinem Bruder Clemens Liegl, die mich durch meine Zeit als Doktorand emotional begleitet haben und stets ein offenes Ohr für mich hatten.

Zu guter Letzt: Dinah, danke für deine Geduld, deine emotionale Unterstützung, die große Wertschätzung, die du mir und meiner Arbeit entgegenbringst, sowie für dein Vertrauen in mich und meine Fähigkeiten. Das alles war beim Verfassen dieser Arbeit unglaublich hilfreich.