ORIGINAL ARTICLE

Language-related differential item functioning between English and German PROMIS Depression items is negligible

H. Felix Fischer^{1,2} | Inka Wahl³ | Sandra Nolte^{1,4} | Gregor Liegl¹ | Elmar Brähler^{5,6} | Bernd Löwe³ | Matthias Rose^{1,7}

¹Department of Psychosomatic Medicine, Clinic for Internal Medicine, Charité - Universitätsmedizin Berlin, Berlin, Germany

²Institute for Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin, Germany

³Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf & Schön Klinik Hamburg Eilbek, Hamburg, Germany

⁴Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Melbourne, Victoria, Australia

⁵Department of Medical Psychology and Medical Sociology, University of Leipzig, Leipzig, Germany

⁶Klinik und Poliklinik für Psychosomatische Medizin und Psychotherapie, Universitätsmedizin Mainz, Mainz, Germany

⁷Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, Massachusetts, USA

Correspondence

Felix Fischer, Institute for Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Luisenstrasse 57, 10117 Berlin, Germany.
Email: felix.fischer@charite.de

Funding information

National Institutes of Health (NIH) Common Fund Initiative, Grant/Award Numbers: U54AR057951, U01AR052177, U54AR057943, U54AR057926, U01AR057948, U01AR052170, U01AR057954, U01AR052171, U01AR052181, U01AR057956, U01AR052158, U01AR057929, U01AR057936, U01AR052155, U01AR057971, U01AR057940, U01AR057967 and U01AR052186.

Abstract

To investigate differential item functioning (DIF) of PROMIS Depression items between US and German samples we compared data from the US PROMIS calibration sample ($n = 780$), a German general population survey ($n = 2,500$) and a German clinical sample ($n = 621$). DIF was assessed in an ordinal logistic regression framework, with 0.02 as criterion for R^2 -change and 0.096 for Raju's non-compensatory DIF. Item parameters were initially fixed to the PROMIS Depression metric; we used plausible values to account for uncertainty in depression estimates. Only four items showed DIF. Accounting for DIF led to negligible effects for the full item bank as well as a *post hoc* simulated computer-adaptive test (< 0.1 point on the PROMIS metric [mean = 50, standard deviation = 10]), while the effect on the short forms was small (< 1 point). The mean depression severity (43.6) in the German general population sample was considerably lower compared to the US reference value of 50. Overall, we found little evidence for language DIF between US and German samples, which could be addressed by either replacing the DIF items by items not showing DIF or by scoring the short form in German samples with the corrected item parameters reported.

KEYWORDS

depression, differential item functioning, Item-Response Theory, outcome assessment, patient-reported outcomes, PROMIS

1 | INTRODUCTION

Over the past years, the Patient-reported Outcomes Measurement Information System (PROMIS) initiative has developed item banks for the measurement of a large number of clinically important outcomes in order to help standardize the measurement of patients' subjective

health status (Reeve et al., 2007; Cella et al., 2007b; Cella et al., 2010b). Item Response Theory (IRT) methods that are generally used for developing and evaluating respective instruments go beyond the concepts of Classical Test Theory (CTT) by modeling the relations between item responses and an underlying latent trait (usually referred to as theta) in a mathematical way (Embretson & Reise, 2000).

Resulting IRT-based item banks are collections of items and their respective parameters assessing a common latent dimension, thereby defining the latent construct being measured. IRT delivers greater flexibility when tailored measures are needed to fit a specific sample, when test scores of different measures need to be compared, and when response data is incomplete (Streiner, 2010; Thomas, 2010; Reise & Waller, 2009; Wahl et al., 2014). Furthermore, IRT models allow computerized adaptive testing (CAT) to increase measurement precision (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Cella et al., 2007a).

As the IRT-based item banking approaches are especially valuable to investigate differences between different language versions of a measure, PROMIS has recently begun to translate its instruments in other languages to further standardize measurement of Patient-Reported Outcomes (PROs) (Alonso et al., 2013). There are differences between cultures regarding constructs such as depression (Ryder et al., 2008) and it has been advised to test psychometric properties of a measure after translation (Beaton, Bombardier, Guillemin, & Ferraz, 2000), as differences in psychometric properties are informative about differential validity of certain items (Thissen, Reeve, Bjorner, & Chang, 2007). Such assessment is particularly important when one wants to compare such different groups on the measure in question (Petersen et al., 2003), since measurement invariance is a prerequisite for valid group comparisons (Meredith & Teresi, 2006).

Based on IRT models it can be evaluated whether items perform similarly across different groups (Hahn, Bode, Du, & Cella, 2006). Differential item functioning (DIF) is a measurement bias between certain groups (e.g. sex, age, language) that leads to systematically different item and, hence, test scores, although the underlying latent variable of interest is constant (Millsap, 2011). There are different approaches to assess DIF based on either factor-analytic methods (for example, multigroup confirmatory factor analysis (CFA) or MIMIC modeling [Finch, 2005]) or IRT methods (for example, IRT-based Likelihood ratio tests [Edelen, Thissen, & Teresi, 2006], Raju's DFIT method [Oshima & Morris, 2008] or in an ordinal regression framework [Crane, Gibbons, Jolley, & Van Belle, 2006]). The assessment of DIF is considered important in validity assessment of PROs, although criteria to detect and estimate impact on scale scores are somewhat unclear (Fayers, 2007). It has been shown that DIF between groups can have considerable impact (Huang, Church, & Katigbak, 1997; Choi et al., 2009; McKenna, 2011), for example, language DIF of various degrees has been found in measures for dementia (Edelen et al., 2006; Crane, Gibbons, Jolley, & Belle, 2006) and quality of life (Hahn et al., 2006; Petersen et al., 2003; Rocha, Power, Bushnell, & Fleck, 2012; Perkins, Stump, Monahan, & McHorney, 2006). Hence, PROMIS considers DIF testing between sex, age, and diagnostic groups as a standard in instrument development (Patient-reported Outcomes Measurement Information System, 2013) and translation (Alonso et al., 2013). Items exhibiting DIF are usually excluded from item banks during development (Pilkonis et al., 2011).

Some PROMIS item banks have already been evaluated regarding language DIF, in particular Physical Functioning between US and Latino (Paz, Spritzer, Morales, & Hays, 2013) as well as US and Dutch samples (Oude Voshaar et al., 2014), Pain Interference between US and Dutch samples (Crins et al., 2015), and Social Health between US and Spanish-speaking samples (Hahn et al., 2014). While a substantial

impact of language related DIF was found in Physical Functioning between US and Latino (Paz et al., 2013), in others language-related DIF was negligible (Crins et al., 2015; Oude Voshaar et al., 2014) or even absent (Hahn et al., 2014). In studies investigating language DIF of depression measures, DIF has been found with various impact on scale scores (Azocar, Areán, Miranda, & Muñoz, 2001; Arthurs, Steele, Hudson, Baron, & Thombs, 2012; Hirsch, Donner-Banzhoff, & Bachmann, 2013; Kwakkenbos et al., 2013) and authors emphasize that assessment of language DIF should be conducted when pooling different language data (Kwakkenbos et al., 2013; Arthurs et al., 2012).

The PROMIS depression item bank, initially calibrated in a US sample of 14,898 participants, outperforming legacy instruments in terms of measurement precision (Pilkonis et al., 2011), was translated into German (Wahl, Löwe, & Rose, 2011). In this paper we investigate language DIF between the US and German PROMIS depression item bank, using data from a general population survey and a clinical sample from Germany as well as US calibration data to assess whether US item parameters are suitable for German-speaking samples. In our DIF analysis, we directly use the PROMIS Depression US parameters as baseline model instead of separate model estimation and linking, as done in earlier works. We also account for uncertainty in theta estimates by using the plausible draws approach (Gorter, Fox, & Twisk, 2015).

2 | METHODS

2.1 | Measures and translation

All 28 items of the PROMIS Depression item bank were translated into German according to state-of-the-art methods and adhering to the PROMIS translation protocol as described elsewhere (Patient-reported Outcomes Measurement Information System, 2013; Wahl et al., 2011). As convergent measures of depression severity, the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001; Martin, Rief, Klaiberg, & Braehler, 2006) was used in the German clinical sample and its respective short form PHQ-2 (Löwe et al., 2010; Kroenke, Spitzer, Williams, & Löwe, 2009) in the German general population sample. In contrast in the US, the Center for Epidemiologic Studies Depression Scale (CES-D) was collected in PROMIS wave 1 (Cella et al., 2010a).

2.2 | Samples

As US sample we used the records from the PROMIS wave 1 data, which were collected through an independent polling company and used for initial calibration of the PROMIS measures (Cella et al., 2010a). Quota sampling was conducted in order to resemble the marginal distributions of gender, age, education and ethnicity (Cella et al., 2010a). Participants fulfilling the following criteria were included in our analysis:

- had all PROMIS depression items presented
- at least 50% of the 28 final depression items were answered
- did not meet the PROMIS exclusion criteria (average response time < 1 second or 10 consecutive items with response time < 0.5 seconds)

We excluded respondents from the block design used for the PROMIS wave 1 data collection (Pilkonis et al., 2011) as these respondents often had answered less than six items of the final depression item bank. These short response patterns resulted in theta estimates with high standard error. Overall, 780 respondents from the US sample fulfilled the above mentioned criteria. Since those respondents are a subsample of the PROMIS scale setting sample, the sample is from the general population but not representative.

The translated PROMIS Depression item bank was presented to a randomly (random-route) generated sample of the German general population. Overall, the sample comprised 4,455 persons of whom 2,504 answered a structured questionnaire personally in the presence of an interviewer. Main reasons for non-participation were refusal to take part (28.4% of 4,455 persons) and not encountered at home (13.9%) (Häuser, Schmutzer, Brähler, & Glaesmer, 2011). This survey was carried out by an independent social research institute (USUMA, Berlin). More detailed information on the sampling scheme can be found elsewhere (Häuser et al., 2011).

A clinical convenience sample ($n = 643$ patients) of the Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, with a wide range of somatic and mental conditions, answered these items during psychometric routine diagnostics. Data from the German general population and clinical samples were pooled to achieve coverage over the whole continuum of depression severity.

The earlier mentioned criterion on missing item responses was applied to the German general population and clinical sample, leading to the exclusion of 4 and 22 records, respectively. The final sample sizes were $n = 2,500$ and $n = 621$ persons, respectively.

2.3 | Statistical analysis

2.3.1 | Depression mean scores from legacy and PROMIS measures

To allow comparison between the legacy measures PHQ-9, PHQ-2 and CES-D, we scaled them to the US PROMIS metric, using item parameters reported by Choi, Schalet, Cook, and Cella (2014). For this, we used our new web application (<http://www.common-metrics.org>). We drew 25 sets of plausible values (Gortler et al., 2015) from the resulting expected a posteriori (EAP) estimates and estimated the mean for each sample in a linear model for each draw. The mean estimates from these 25 models were then combined according to Rubin's rule (Schafer & Graham, 2002). We calculated the depression severity from the PROMIS item bank, using US item parameters (Pilkonis et al., 2011), accordingly.

2.3.2 | Unidimensionality and local independence

To assess unidimensionality of the data, a CFA model was fitted using the weighted least squares means and variance adjusted (WLSMV) estimator to account for the ordinal nature of the responses. Since for a unidimensional model insufficient fit has been reported in a German sample, we also fitted the proposed bifactor model (Jakob et al., 2015). We assessed chi-square, the Comparative Fit Index (CFI, cutoff >0.95), the Tucker-Lewis Index (TLI, cutoff >0.95) and the root mean square error of approximation (RMSEA, cutoff <0.08)

as measures of model fit (Brown & Kenny, 2006; Hu & Bentler, 1999). Furthermore, we investigated size of residual correlations to identify possible violations of local independence and estimated exploratory bifactor models for each subsample to calculate the explained common variance (ECV, cutoff >0.60) (Reise, Scheines, Widaman, & Haviland, 2012).

2.3.3 | DIF analysis

Items showing DIF regarding sex, age and education during the construction process of the PROMIS Depression item bank were already excluded from final calibration in the US sample (Pilkonis et al., 2011). In German samples DIF of the PROMIS Depression item bank regarding sample origin, age, gender, and level of education was not found in an earlier analysis (Wahl et al., 2014). Hence, we focused on analysis of language-related DIF. Given the rigorous translation process (Patient-reported Outcomes Measurement Information System, 2013; Wahl et al., 2011), we had no specific hypothesis which items might be prone to DIF.

For the assessment of potential language DIF in PROMIS measures, to date researchers have used DIF analysis in an ordinal regression framework (Crane et al., 2006) and its implementation in the lordif-package (Choi, Gibbons, & Crane, 2011) with a criterion of R^2 -change of 0.02 (Crins et al., 2015; Hahn et al., 2014; Oude Voshaar et al., 2014; Paz et al., 2013). In case of absence of language DIF, parameters from the US calibration sample have been simply used in other languages as well (Hahn et al., 2014; Paz et al., 2013). However, in earlier studies DIF analyses included separate model estimation steps in the population of interest and subsequent linking of parameters to the reference metric. In contrast, we used the item parameters reported by Pilkonis et al. (2011) as the baseline model with unconstrained sample mean and standard deviation.

For all items we unconstrained item parameters from the baseline model one at a time in the German sample. The resulting models were used to estimate theta with the EAP technique in both the German and the US sample. Since direct use of EAP estimates in regression models does not account for uncertainty in theta, we drew 25 samples of plausible values (Gortler et al., 2015) and fitted three ordered logistic regression models with the following assumptions in each of the 25 imputed datasets:

- Model 1 (no DIF): item response $\sim$ theta
- Model 2 (uniform DIF): item response $\sim$ theta + language
- Model 3 (non-uniform DIF, interaction term included): item response $\sim$ theta + language + theta*language

After combining R^2 over the 25 datasets (Harel, 2009), items exhibiting an increase in Nagelkerke's pseudo- R^2 by 0.02 between those models were flagged as potential DIF items (Watt et al., 2014). We observed that although the expected test score curves between German and US samples differed to a somewhat meaningful extent but did not meet the criterion of R^2 change >0.02 . Hence, we decided to include Raju's NCDIF (Raju, van der Linden, & Fleer, 1995) with a cutoff of 0.096 as used in earlier studies (Teresi et al., 2007; Deng,

Anatchkova, Waring, Han, & Ware, 2015) as a criterion to flag items for DIF.

Item parameters for the flagged items were then re-estimated in the German sample with latent variable mean and variance constrained to estimates based on the set of anchor items not flagged for DIF. Based on this model, we re-estimated theta, drew 25 sets of plausible values and fitted the three regression models again to flag items for DIF. This was repeated until the same set of items was flagged for DIF in two subsequent runs.

This process resulted in a multigroup GRM with item parameters corrected for language-related DIF in German samples, but still on the same scale as the US parameters. We compared expected test scores between the US model and the DIF corrected model to assess impact of DIF on scale scores.

Using the model corrected for DIF, we estimated 25 sets of plausible values of theta for each respondent and each test form (full bank using all 28 items, the PROMIS Depression short forms 8a, 8b, 6a, 4a as well as CAT). We used Firestar (Choi, 2009) to simulate CAT which was designed to apply 4–12 items until a standard error of 0.32 (corresponding to a reliability of 0.90) is achieved. Precision over the theta continuum was compared. We also investigated the differences found between the individual plausible scores from both models descriptively with loess smoothers. In order to assess a possible mean difference and an effect of depression severity we fitted a linear model; the estimates from the 25 draws were again combined according to Rubin's rule (Lumley, 2014).

We did not consider sampling weights to approximate population representativeness of the German general population sample in our analysis.

All statistical analyses were conducted using R 3.2.2 (R Development Core Team, 2008). In particular, we used packages rms for ordinal

logistic regression analysis (Harrell, 2013), DFIT for NCDIF calculation (Cervantes, 2015), mirt for multiple group IRT model estimation and theta calculation (Chalmers, 2012), mitools to combine estimates (Lumley, 2014) and ggplot2 for graphics (Wickham, 2009).

3 | RESULTS

3.1 | Descriptives

Descriptive sample characteristics are presented in Table 1. While sex, age and marital status are similar between the German general population and the PROMIS US sample, participants from the German clinical sample are considerably younger and more likely to be female. Furthermore, in the US sample low education levels were observed less likely.

Depression severity, as measured with legacy instruments and PROMIS Depression, yields reasonable agreement between the US and the German clinical sample. As expected the clinical sample shows elevated depression severity regardless of the measure used. A large difference of theta estimates is found in the German general population sample, which can be explained by the strong prior influence when estimating theta from a very short measure like the PHQ-2.

3.2 | Unidimensionality

In the unidimensional model all items correlated strongly (0.80 to 0.94) with a single factor. The model yielded reasonable fit (chi-square = 5798.9, df = 350, $p < 0.001$; CFI = 0.985; TLI = 0.984, RMSEA = 0.072 (confidence interval (CI) 0.071, 0.074)); the largest observed residual correlation was 0.10 between eddep27 and eddep28 indicating negligible local dependence. The bifactor model yielded significantly better fit (chi-square = 3354.4, df = 321, $p < 0.001$, scaled chi-

TABLE 1 Descriptives

<i>n</i>	German general population 2500	German clinical population 621	US PROMIS wave 1 780
Age (mean (standard deviation))	50.63 (18.54)	40.45 (14.24)	51.20 (18.98)
Sex = male (%)	1170 (46.8)	185 (29.9)	376 (48.3)
Education ^a (%)			
low	1139 (47.0)	113 (19.5)	8 (1.0)
medium	924 (38.2)	199 (34.3)	165 (21.2)
high	358 (14.8)	268 (46.2)	607 (77.8)
Marital status (%)			
not married	582 (23.3)	301 (49.8)	180 (23.1)
married	1346 (53.8)	200 (33.1)	442 (56.8)
separated, divorced or widowed	572 (22.9)	104 (17.2)	156 (20.1)
Legacy measure theta ^b (mean (CI))	47.9 (47.4; 48.4)	63.4 (62.6; 64.2)	48.7 (47.9; 49.4)
PROMIS Depression theta ^c (mean (CI))	43.6 (43.2; 44.0)	61.2 (60.4; 62.0)	48.5 (47.8; 49.2)

^aUS education data was recoded as follows: fifth grade or less, sixth grade, seventh grade, eighth grade = low, Some high school, High school grad/GED = medium, Some college/Technical degree/AA, College degree (BA/BS), Advanced degree (MA, PHD, MD) = high; German education data: no degree or Hauptschulabschluss = low, Realschulabschluss = medium, Abitur = high.

^bDepression severity was scaled to the US PROMIS metric with item parameters from Choi et al. (2014) and unconstrained mean and standard deviation, using PHQ-2 in the German general population, PHQ-9 in the German clinical population and CES-D in the US PROMIS wave 1 data. For each sample 25 sets of plausible values were drawn, means were estimated in a linear model in each draw, and combined according to Rubin's rule.

^cEstimated using the US item parameters. For each sample 25 sets of plausible values were drawn, means were estimated in a linear model in each draw, and combined according to Rubin's rule.

TABLE 2 Results from DIF analysis

Item		Model 1	Nagelkerke R^2		Uniform	ΔR^2	
			Model 2	Model 3		Non-uniform	NCDIF
eddep04	I felt worthless	0.841	0.845	0.846	0.004	0.005	0.032
eddep05	I felt that I had nothing to look forward to	0.860	0.867	0.867	0.007	0.008	0.103 ^a
eddep06	I felt helpless	0.852	0.860	0.860	0.008	0.008	0.103 ^a
eddep07	I withdrew from other people	0.814	0.814	0.814	<0.001	<0.001	0.018
eddep09	I felt that nothing could cheer me up	0.866	0.872	0.872	0.006	0.006	0.071
eddep14	I felt that I was not as good as other people	0.808	0.809	0.809	0.001	0.001	0.029
eddep17	I felt sad	0.863	0.864	0.864	<0.001	0.001	0.012
eddep19	I felt that I wanted to give up on everything	0.811	0.811	0.812	<0.001	<0.001	0.017
eddep21	I felt that I was to blame for things	0.784	0.789	0.790	0.006	0.006	0.039
eddep22	I felt like a failure	0.831	0.835	0.835	0.005	0.005	0.055
eddep23	I had trouble feeling close to people	0.806	0.811	0.811	0.004	0.004	0.019
eddep26	I felt disappointed in myself	0.819	0.829	0.829	0.011	0.011	0.133 ^a
eddep27	I felt that I was not needed	0.755	0.755	0.759	<0.001	0.004	0.022
eddep28	I felt lonely	0.776	0.777	0.777	<0.001	0.001	0.010
eddep29	I felt depressed	0.871	0.876	0.876	0.004	0.004	0.066
eddep30	I had trouble making decisions	0.799	0.799	0.800	<0.001	<0.001	0.019
eddep31	I felt discouraged about the future	0.846	0.848	0.849	0.002	0.003	0.037
eddep35	I found that things in my life were overwhelming	0.828	0.830	0.831	0.002	0.003	0.005
eddep36	I felt unhappy	0.887	0.890	0.891	0.003	0.004	0.089
eddep39	I felt I had no reason for living	0.731	0.731	0.732	<0.001	0.001	0.011
eddep41	I felt hopeless	0.861	0.862	0.862	<0.001	<0.001	0.025
eddep42	I felt ignored by people	0.733	0.739	0.739	0.006	0.006	0.015
eddep44	I felt upset for no reason	0.743	0.744	0.745	0.001	0.002	0.007
eddep45	I felt that nothing was interesting	0.779	0.779	0.780	0.001	0.001	0.013
eddep46	I felt pessimistic	0.823	0.824	0.824	0.001	0.001	0.031
eddep48	I felt that my life was empty	0.830	0.830	0.830	<0.001	<0.001	0.008
eddep50	I felt guilty	0.739	0.753	0.753	0.014	0.014	0.145 ^a
eddep54	I felt emotionally exhausted	0.841	0.841	0.842	<0.001	0.001	0.034

^aDIF criteria exceeded.

square difference test: $\Delta\chi^2 = 968.7$, $\Delta df = 16.1$, $p < 0.001$, CFI = 0.992, TLI = 0.990, RMSEA = 0.056 (CI 0.055, 0.058)), but the two subfactors correlated strongly ($r = 0.91$), possibly indicating overfitting (Xiong et al., 2014). The explained variance of the general factor in an exploratory bifactor model was 0.91, which is also well above the cutoff of 0.60. Taken together these results strongly suggested suitability for unidimensional IRT analysis.

3.3 | DIF analysis

The first and the second run of the DIF analysis (see Table 2) had the same result: no item showed uniform or non-uniform DIF exceeding the criterion of R^2 change >0.02 ; a total of four items exhibited the NCDIF cutoff of 0.096 (eddep05: "I felt that I had nothing to look forward to", eddep06: "I felt helpless", eddep26: "I felt disappointed in myself" and eddep50: "I felt guilty").

Items eddep05 and eddep06 were scored higher in German samples assuming constant depression severity, while items eddep26 and eddep50 were scored lower. Specific GRM parameters of these four items for use in German samples can be found in Table 3. Correcting for DIF led to a maximum difference in expected test scores of 0.27

(full bank, given a possible maximum score of 112), 0.95 (short forms 8a, 8b, maximum score 32), and 0.47 (short form 4a, maximum score 16, short form 6a, maximum score 24).

The CAT applied an average of 7.7 (standard deviation = 3.7) items in the German general population and 4.4 (1.4) items in the German clinical sample. In 37% ($n = 928$) of the general population sample the CAT did not reach the precision criterion, which was true in only 2% (15) of the cases in the clinical sample. As expected longer test forms result in smaller standard errors of theta estimates, but it is noteworthy that even the shortest measure (short form 4a) reaches a reliability of about 0.9 when theta is between 50 and 70. The CAT results in the most constant measurement precision over the whole continuum.

TABLE 3 Item parameters for the items showing DIF re-estimated in the German sample

Item	a	b1	b2	b3	b4
eddep05	3.455	-0.190	0.564	1.336	2.352
eddep06	3.408	-0.049	0.583	1.269	2.299
eddep26	2.819	0.014	0.829	1.491	2.469
eddep50	2.299	0.568	1.301	2.011	2.869

FIGURE 1 Bland-Altman plots comparing theta estimates from the US and the DIF corrected model for the full itembank, a CAT and the PROMIS Depression short forms. For each set of plausible values a loess smoother shows the mean agreement. Histogram shows distribution of German general population and clinical sample. Overall, mean theta difference is small over the continuum

TABLE 4 Difference of depression estimates from US and DIF corrected model was predicted by depression severity, regression estimates from a linear model with 95% confidence interval are shown.

Form	Intercept	mean theta (from US and DIF corrected model)
Full Bank	0.03 [−0.15; 0.21]	−0.01 [−0.03; 0.00]
CAT	−0.09 [−0.33; 0.15]	−0.02 [−0.05; 0.00]
Short form 8a	−0.98 [−1.31; −0.64]	0.02 [−0.01; 0.05]
Short form 8b	−0.76 [−1.03; −0.48]	0.00 [−0.02; 0.03]
Short form 6a	−0.54 [−0.84; −0.24]	0.01 [−0.02; 0.03]
Short form 4a	−0.93 [−1.22; −0.64]	0.03 [0.00; 0.06]

Bland-Altman plots (Figure 1) revealed that over all sets of plausible values, the differences between the US model and the DIF corrected model were not linear, but deviations from linearity were mostly at extreme values of depression severity. The DIF corrected model resulted in larger depression estimates in the CAT in the clinical sample, whereas the short form 8a and short form 4a yielded larger depression estimates in the general population. Overall, these effects appeared to be small. Table 4 shows the effect of DIF correction on theta estimates from the full bank, a CAT and the short forms over 25 sets of plausible values when assuming a linear effect. Overall, correction for DIF had a negligible impact on scores on the group level for the full bank and the CAT, with the overall mean of scores differing less than 0.1 points. For the short forms, the mean difference was larger and statistically significant, but still not exceeding a small effect size on the T-metric (< 1 point). In all forms, we only found small effects of depression severity on the difference of theta estimates.

4 | DISCUSSION

We found that the German translation of the PROMIS Depression item bank is a unidimensional and precise measure for depression. No item showed language-related DIF in German samples compared to the US item parameters according to the common criterion of DIF in an ordinal logistic regression framework as used in other translations of PROMIS measures before. Further, using a more conservative criterion as Rajus NCDIF (Raju et al., 1995), only 4 out of 28 items showed language DIF. Accounting for these differences led to only minor differences in depression estimates in all short forms. Taken together, this finding provides strong evidence that use of US parameters is warranted in German samples, unless one wants to investigate small effects. In that case, it might be necessary to take DIF into account. Unfortunately, all current short forms made available by PROMIS (4a, 6a, 8a, 8b) contain at least one item exhibiting DIF in our analysis. This could be addressed by either replacing the DIF items by items not showing DIF

with similar IRT parameters or by scoring the short form in German samples with the corrected item parameters reported in this paper.

We observed a large mean difference between the German and the US general population samples which cannot be explained by DIF. Both samples deviated by more than half a standard deviation on the T-score metric. The difference in depression severity could result from differences between populations; however, recent national survey data suggest a strikingly similar 12-month prevalence estimate for major depression, i.e. 6.0% in Germany (Jacobi et al., 2014) and 6.6% in the US, respectively (Center for Behavioral Health Statistics and Quality 2015).

Both samples might differ in other variables beyond depressive severity, e.g. because of the different sampling strategies employed or through selective non-participation. For example, the German general population sample has a higher percentage of females and participants with lower educational levels compared to the Mikrozensus 2014. This would rather point to an increased depression severity in the German general population sample. Lack of comparable socio-demographic data made investigation of this issue difficult. It may therefore be valuable that internationally comparable measures of socio-demographic data be used in studies contributing to the development of a common PROMIS metric.

Furthermore, data were collected slightly differently between samples. The German general population sample answered the items on a paper questionnaire, while the data from the German clinical sample and the PROMIS US sample were obtained electronically (via handhelds or via internet). However, such different modes of questionnaire administration have been shown to have little effect (Bjorner et al., 2013). We observed a stronger floor effect in the German general population than in the US general population, with about one quarter of the German general population sample answering “never”, i.e. the lowest response category, to every single item presented. The presence of an interviewer in the German general population sample might have forced respondents to answer in a more socially acceptable way. Such an influence on certain items should be detectable by our DIF analysis, whereas a general dissimulation of depressive symptoms could not be detected. A further possibility for the considerable mean difference in depression severity could be that the interpretation of the response format differs between languages. These issues could unfortunately not be addressed within the current study design.

The strengths of our study are that we were able to use a large sample from the German general population combined with a clinical sample of patients with elevated depression scores. Furthermore, we based our DIF analysis on the published and widely used US parameters (Pilkonis et al., 2011) which made re-estimation of item parameters in the US sample unnecessary. We also accounted for the uncertainty of theta estimates by using plausible values instead of the crude EAP estimates throughout the paper (Gortler et al., 2015). This ensures that the less precise EAP estimates have less impact on the analysis as the more precise EAP estimates. Taking this uncertainty into account, the differences in R^2 between the different regression models in the DIF analysis were smaller than when based on crude EAP estimates.

A crucial limitation of our analysis is the selection of a somewhat arbitrary DIF criterion, but unfortunately, there is no consensus which criteria for DIF detection balance power and clinical usefulness.

However, R^2 has been used as a standard measure in PROMIS, but has also been reported to be potentially not sensitive enough to identify meaningful DIF (Crane et al., 2007). In contrast, Likelihood ratio tests between IRT models are reported to be too sensitive to deviations from the sample-specific item parameters (Edelen, Stucky, & Chandra, 2015). Hence, different cutoffs of R^2 -change for relevant DIF have been used, e.g. 0.02 (Watt et al., 2014), 0.03 (Wahl et al., 2014); Fliege et al., 2005) or 0.035 and higher (Walter et al., 2007; Resnik, Tian, Ni, & Jette, 2012). Since R^2 might inflate in homogenous samples, it is problematic to compare R^2 -change over studies. Within PROMIS there is consensus to use a rather conservative R^2 -change criterion and we also included Najus NCDIF (Raju et al., 1995). As an exploratory sensitivity analysis based on a reviewer's suggestion, we calculated the area under the curve (AUC, c-statistic) and its change between models. In general, AUC was high for each item (0.87 to 0.95) and its DIF-related change very small (0.00 to 0.01), which is consistent with our interpretation of R^2 -change. Still, it appears necessary to investigate new measures of DIF (Edelen et al., 2015).

While we had some information on the reasons for non-participation in the German general population sample, we did not have any information on non-participation in the US and the German clinical sample. We believe an influence of non-participation on our study is unlikely, since a possible selection would be probably associated with age, gender, education or health status; it has been shown before that none of the items in question showed DIF related to these variables neither in English nor German samples.

Within this paper we assessed scale equivalence, but establishment of metric equivalence remains unsolved (Bullinger, Anderson, Cella, & Aaronson, 1993). We have shown that a theta of 50 – indicating the mean depressive severity of the US calibration sample – does not resemble the mean depressive severity in the German general population sample. We feel that there are two competing viable approaches to address such differences between two language versions of a test, with both having their advantages and shortcomings. One approach would be to use the same item parameters for every language group. Item parameters could be corrected for DIF, if necessary. This approach allows cross-language comparisons of individuals and the same response patterns would result in similar theta estimates in all groups. However, the scale would have a rather arbitrary mean and standard deviation in translated versions, which depends on the calibration sample in the original version.

Another approach aims to achieve that for all language versions of a given instrument representative general population sample score with a mean of 50 and a standard deviation of 10. This leads to a standard and widely valid definition of any metric by providing a strong and language-specific anchor calibrated to the respective representative general population sample; but at the same time, it leads to problems in cross-language studies where two persons' latent trait may have the same value but different meanings due to overall differences in trait distributions between general population samples representing different language and cultural groups. Some sort of relative comparability (relative to the respective representative sample) could be established using linking-methods such as equipercentile-linking (Dorans, 2007), but it would be a huge effort to have comparable, yet representative general population samples in each language.

Given that PROMIS International aims to establish translations in more than two languages, this is an important issue. The argument can be made that a common metric will be more likely accepted as a standard, when item parameters are globally the same. Hence, for an internationally comparable outcome measurement, one long-term perspective might be to aim for a global reference point on the scale.

ACKNOWLEDGMENTS

PROMIS was funded with cooperative agreements from the National Institutes of Health (NIH) Common Fund Initiative (U54AR057951, U01AR052177, U54AR057943, U54AR057926, U01AR057948, U01AR052170, U01AR057954, U01AR052171, U01AR052181, U01AR057956, U01AR052158, U01AR057929, U01AR057936, U01AR052155, U01AR057971, U01AR057940, U01AR057967, U01AR052186). The contents of this article uses data developed under PROMIS. These contents do not necessarily represent an endorsement by the US Federal Government or PROMIS. See www.nihpromis.org for additional information on the PROMIS initiative.

Declaration of Interest Statement

The authors have no competing interests.

REFERENCES

- Alonso, J., Bartlett, S. J., Rose, M., Aaronson, N. K., Chaplin, J. E., Efficace, F., ... Forrest, C. B. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS) initiative. *Health and Quality of Life Outcomes*, 11(210), 1–5. doi:10.1186/1477-7525-11-210
- Arthurs, E., Steele, R. J., Hudson, M., Baron, M., & Thombs, B. D. (2012). Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. *PLoS One*, 7(12), e52028. doi:10.1371/journal.pone.0052028
- Azocar, F., Areán, P., Miranda, J., & Muñoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology*, 57(3), 355–365.
- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186–3191.
- Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. (2013). Difference in method of administration did not significantly impact item response: an IRT-based analysis from the patient-reported outcomes measurement information system (PROMIS) initiative. *Quality of Life Research*. doi:10.1007/s11136-013-0451-4
- Brown, T. A., & Kenny, D. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.
- Bullinger, M., Anderson, R., Cella, D., & Aaronson, N. K. (1993). Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Quality of Life Research*, 2(6), 451–459. doi:10.1007/BF00422219
- Cella, D., Gershon, R., Lai, J.-S., Choi, S. W., Yount, S., Rothrock, N., ... Rose, M. (2007a). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, 16(5 Suppl 1), 133–141. doi:10.1007/s11136-007-9204-6
- Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., ... Rose, M. (2007b). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care*, 45(5), 3–11. doi:10.1097/01.mlr.0000258615.42478.55
- Cella, D., Riley, W., Stone, A., Northrock, N., Reeve, B. B., Yount, S., ... Hays, R. (2010a). Initial Adult Health Item Banks and First Wave Testing of the Patient-Reported Outcomes Measurement Information System (PROMIS™) Network: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194. doi:10.1016/j.jclinepi.2010.04.011.Initial.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B. B., Yount, S., ... Hays, R. (2010b). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194. doi:10.1016/j.jclinepi.2010.04.011.
- Center for Behavioral Health Statistics and Quality, (2015). Behavioral health trends in the United States: Results from the 2014 National Survey on Drug Use and Health. HHS Publication No. SMA 15-4927, NSDUH Series H-50. Available at: <http://www.samhsa.gov/data/>.
- Cervantes, V. H. (2015). *DFIT: An R package for the Differential Functioning of Items and Tests framework*.
- Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. doi:10.18637/jss.v048.i06
- Choi, S. W. (2009). FIRESTAR : Computerized Adaptive Testing (CAT) Simulation Program for Polytomous IRT models. <http://www.nihpromis.org/resources/resourcehome>
- Choi, B., Bjorner, J. B., Ostergren, P.-O., Clays, E., Houtman, I., Punnett, L., ... Karasek, R. (2009). Cross-language differential item functioning of the job content questionnaire among European countries: The JACE study. *International Journal of Behavioral Medicine*, 16(2), 136–147. doi:10.1007/s12529-009-9048-2
- Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, 19(1), 125–136. doi:10.1007/s11136-009-9560-5
- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39(8), 1–30. doi:10.18637/jss.v039.i08
- Choi, S. W., Schalet, B. D., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment*, 26(2), 513–527. doi:10.1037/a0035768
- Crane, P. K., Gibbons, L. E., Jolley, L., & Belle, G. V. (2006). Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques. *Medical Care*, 44(11), 115–123. doi:10.1097/01.mlr.0000245183.28384.ed
- Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., ... Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, 16(Suppl 1), 69–84. doi:10.1007/s11136-007-9185-5
- Crins, M. H. P., Roorda, L. D., Smits, N., de Vet, H. C. W., Westhovens, R., Cella, D., ... Terwee, C. B. (2015). Calibration and validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. *PLoS One*, 10(7), e0134094. doi:10.1371/journal.pone.0134094
- Deng, N., Anatchkova, M. D., Waring, M. E., Han, K. T., & Ware, J. E. (2015). Testing item response theory invariance of the standardized Quality-of-life Disease Impact Scale (QDIS®) in acute coronary syndrome patients: Differential functioning of items and test. *Quality of Life Research*, 24(8), 1809–1822. doi:10.1007/s11136-015-0916-8
- Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16(Suppl 1), 85–94. doi:10.1007/s11136-006-9155-3
- Edelen, M. O., Thissen, D., & Teresi, J. A. (2006). Identification of Differential Item Functioning using Item Response Theory and the Likelihood-based model comparison approach: Application to the Mini-Mental

- State Examination. *Medical Care*, 44(11), 134–142. doi:10.1097/01.mlr.0000245251.83359.8c
- Edelen, M. O., Stucky, B. D., & Chandra, A. (2015). Quantifying “problematic” DIF within an IRT framework: Application to a cancer stigma index. *Quality of Life Research*, 24(1), 95–103. doi:10.1007/s11136-013-0540-4
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research*, 16(Suppl 1), 187–194. doi:10.1007/s11136-007-9197-1
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel–Haenszel, SIBTEST, and the IRT Likelihood ratio. *Applied Psychological Measurement*, 29(4), 278–295. doi:10.1177/0146621605275728
- Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, 14(10), 2277–2291. doi:10.1007/s11136-005-6651-9
- Gortler, R., Fox, J.-P., & Twisk, J. (2015). Why Item Response Theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, 2, 1–12. doi:10.1186/s12874-015-0050-x
- Hahn, E. A., Bode, R. K., Du, H., & Cella, D. (2006). Evaluating linguistic equivalence of patient-reported outcomes in a cancer clinical trial. *Clinical Trials*, 3(3), 280–290. doi:10.1191/1740774506cn1480a
- Hahn, E. A., DeWalt, D. A., Bode, R. K., Garcia, S. F., Devellis, R. F., Correia, H., ... Cella, D. (2014). New English and Spanish social health measures will facilitate evaluating health determinants. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, 33(5), 490–499. doi:10.1037/hea0000055
- Harel, O. (2009). The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, 36(10), 1109–1118. doi:10.1080/02664760802553000
- Harrell, F. E. (2013). *rms: Regression Modeling Strategies*. Nashville, TN: Department of Biostatistics, Vanderbilt University.
- Häuser, W., Schmutzer, G., Brähler, E., & Glaesmer, H. (2011). Maltreatment in childhood and adolescence: results from a survey of a representative sample of the German population. *Deutsches Ärzteblatt International*, 108(17), 287–294. doi:10.3238/arztebl.2011.0287
- Hirsch, O., Donner-Banzhoff, N., & Bachmann, V. (2013). Measurement equivalence of four psychological questionnaires in native-born Germans, Russian-speaking immigrants, and native-born Russians. *Journal of Transcultural Nursing: Official Journal of the Transcultural Nursing Society/Transcultural Nursing Society*, 24(3), 225–235. doi:10.1177/1043659613482003
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. doi:10.1080/10705519909540118
- Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential Item Functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, 28(2), 192–218. doi:10.1177/0022022197282004
- Jacobi, F., Höfler, M., Siegert, J., Mack, S., Gerschler, A., Scholl, L., ... Wittchen, H. U. (2014). Twelve-month prevalence, comorbidity and correlates of mental disorders in Germany: the Mental Health Module of the German Health Interview and Examination Survey for Adults (DEGS1-MH). *International Journal of Methods in Psychiatric Research*, 23(3), pp. 304–319. doi:10.1002/mpi
- Jakob, T., Nagl, M., Gramm, L., Heyduck, K., Farin, E., & Glattacker, M. (2015). Psychometric properties of a German translation of the PROMIS(R) Depression item bank. *Evaluation & the Health Professions*, 1–15. doi:10.1177/0163278715598600
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613. doi:10.1046/j.1525-1497.2001.016009606.x
- Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ-4. *Psychosomatics*, 50(6), 613–621. doi:10.1016/S0033-3182(09)70864-3
- Kwakkenbos, L., Arthurs, E., van den Hoogen, F. H. J., Hudson, M., van Lankveld, W. G. J. M., Baron, M., ... Thombs, B. D. (2013). Cross-language measurement equivalence of the Center for Epidemiologic Studies Depression (CES-D) Scale in Systemic Sclerosis: A Comparison of Canadian and Dutch Patients. *PloS One*, 8(1) e53923. doi:10.1371/journal.pone.0053923
- Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., ... Brähler, E. (2010). A 4-item measure of depression and anxiety: validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders*, 122(1–2), 86–95. doi:10.1016/j.jad.2009.06.019
- Lumley, T. (2014). *mitools: Tools for multiple imputation of missing data*.
- Martin, A., Rief, W., Klaiberg, A., & Braehler, E. (2006). Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *General Hospital Psychiatry*, 28(1), 71–77. doi:10.1016/j.genhosppsych.2005.07.003
- McKenna, S. P. (2011). Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. *BMC Medicine*, 9(1), 86. doi:10.1186/1741-7015-9-86
- Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44(11 Suppl 3), S69–S77. doi:10.1097/01.mlr.0000245438.73837.89
- Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.
- Oshima, T. C., & Morris, S. B. (2008). Raju's Differential Functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, 27(3), 43–50. doi:10.1111/j.1745-3992.2008.00127.x
- Oude Voshaar, M. A. H., Ten Klooster, P. M., Glas, C. A. W., Vonkeman, H. E., Taal, E., Krishnan, E., ... van de Laar, M. A. F. J. (2014). Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. *PloS One*, 9(3) e92367. doi:10.1371/journal.pone.0092367
- Patient-reported Outcomes Measurement Information System. (2013). PROMIS Instrument Development and Validation Scientific Standards Version 2.0., 0(May). http://www.nihpromis.org/Documents/PROMISStandards_Vers2.0_Final.pdf [20 March 2016].
- Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Quality of Life Research*, 22(7), 1819–1830. doi:10.1007/s11136-012-0292-6
- Perkins, A. J., Stump, T. E., Monahan, P. O., & McHorney, C. A. (2006). Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Quality of Life Research*, 15(3), 331–348. doi:10.1007/s11136-005-1551-6
- Petersen, M., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., ... Sullivan, M. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research*, 12(4), 373–385. doi:10.1023/A:1023488915557
- Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, 18(3), 263–283. doi:10.1177/1073191111411667
- R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of Differential Functioning of items and tests. *Applied Psychological Measurement*, 19(4), 353–368. doi:10.1177/014662169501900405

- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45 (5 Suppl 1), S22–S31. doi:10.1097/01.mlr.0000250483.85507.04
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27–48. doi:10.1146/annurev.clinpsy.032408.153553
- Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5–26. doi:10.1177/0013164412449831
- Resnik, L., Tian, F., Ni, P., & Jette, A. (2012). Computer-adaptive test to measure community reintegration of Veterans. *Journal of Rehabilitation Research and Development*, 49(4), 557–566.
- Rocha, N. S., Power, M. J., Bushnell, D. M., & Fleck, M. P. (2012). Cross-cultural evaluation of the WHOQOL-BREF domains in primary care depressed patients using Rasch analysis. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 32(1), 41–55. doi:10.1177/0272989X11415112
- Ryder, A. G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S. J., ... Bagby, R. M. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, 117(2), 300–313. doi:10.1037/0021-843X.117.2.300
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. doi:10.1037//1082-989X.7.2.147
- Streiner, D. L. (2010). Measure for measure: new developments in measurement and item response theory. *Canadian Journal of Psychiatry*, 55 (3), 180–186.
- Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., ... Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research*, 16(Suppl 1), 43–68. doi:10.1007/s11136-007-9186-4
- Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, 16(Suppl 1), 109–119. doi:10.1007/s11136-007-9169-5
- Thomas, M. L. (2010). The value of item response theory in clinical assessment: A review. *Assessment*, 18(3), 291–307. doi:10.1177/1073191110374797
- Wahl, I., Löwe, B., & Rose, M. (2011). Das Patient-Reported Outcomes Measurement Information System (PROMIS): Übersetzung der Item-Banken für Depressivität und Angst ins Deutsche. *Klinische Diagnostik und Evaluation*, 4, 236–261.
- Wahl, I., Löwe, B., Bjorner, J. B., Fischer, H. F., Langs, G., Voderholzer, U., ... Rose, M. (2014). Standardization of depression measurement: A common metric was developed for 11 self-report depression measures. *Journal of Clinical Epidemiology*, 67(1), 73–86. doi:10.1016/j.jclinepi.2013.04.019
- Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for “Anxiety” (Anxiety-CAT). *Quality of Life Research*, 16(Suppl 1), 143–155. doi:10.1007/s11136-007-9191-7
- Watt, T., Groenvold, M., Hegedüs, L., Bonnema, S. J., Rasmussen, A. K., Feldt-Rasmussen, U., ... Bjorner, J. B. (2014). Few items in the thyroid-related quality of life instrument ThyPRO exhibited differential item functioning. *Quality of Life Research*, 23(1), 327–338. doi:10.1007/s11136-013-0462-1
- Wickham, H. (2009). *ggplot2*. New York: Springer.
- Xiong, N., Fritzsche, K., Wei, J., Hong, X., Leonhart, R., Zhao, X., ... Fischer, H. F. (2014). Validation of patient health questionnaire (PHQ) for major depression in Chinese outpatients with multiple somatic symptoms: A multicenter cross-sectional study. *Journal of Affective Disorders*, 174, 636–643. doi:10.1016/j.jad.2014.12.042

How to cite this article: Fischer HF, Wahl I, Nolte S, et al. Language-related differential item functioning between English and German PROMIS Depression items is negligible. *Int J Methods Psychiatr Res*. 2017;26:e1530. <https://doi.org/10.1002/mpr.1530>