Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France and Germany

Felix Fischer[1], Chris Gibbons[2], Joel Coste[3], Jose M. Valderas[4], Matthias Rose[1,5]* & Alain Leplege[6]*

*: Both authors contributed equally

1 Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Germany
2 Psychometrics Centre, University of Cambridge, United Kingdom
3 APEMAC, EA 4360, Paris Descartes University and Epidemiology Unit, Hôtel Dieu, Assistance Publique, Hôpitaux de Paris, France
4 University of Exeter, Health Services & Policy Research Group, Exeter, United Kingdom
5 Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester
6 Département d'Histoire et de Philosophie des Sciences, 6 Universite Paris Diderot, France

Corresponding author:
Felix Fischer
Department of Psychosomatic Medicine
Charité Universitätsmedizin Berlin
Charitéplatz 1
D-10117 Berlin
Germany
Tel.: +49/30/450-529 104
Fax: +49/30/450-553 989
e-Mail: felix.fischer@charite.de

Word count: approximately 4,500 words
Figures: 2
Tables: 5

Acknowledgements

This study was funded by the Centre Virchow-Villerme (https://virchowvillerme.eu/). We like to acknowledge the many people involved in development and translation of the PROMIS measures used in this study. Our thanks for their efforts to translate various PROMIS measures into German and French go in particular to Susan Bartlett, Marie-Eve Carrier, Erik Farin-Glattacker, Katja Heyduck, Sandra Nolte and Inka Wahl. We thank Laurence Erdur and Nina Obbarius for their help comparing the different language versions and Terrence Jorgensen for illuminating the pitfalls in measurement invariance testing of ordinal data. Furthermore, we would like to address special thanks to PROMIS translation manager Helena Correia.

Purpose: Comparability of Patient-Reported outcome measures over different languages is essential to allow cross-national research. We investigate the comparability of the PROMIS Profile 29, a generic health-related quality of life measure, in general population samples in the UK, France and Germany and present general population reference values.

Methods: A web based survey was simultaneously conducted in the UK (n = 1,509), France (1,501) and Germany (1,502). Along with the PROMIS profile 29, we collected sociodemographic information as well as the EQ-5D. We tested measurement invariance by means of multi-group confirmatory factor analysis. Differences in the health-related quality of life between countries were modeled by linear regression analysis. We present general population reference data for the included PROMIS domains utilizing plausible value imputation and quantile regression.

Results: Multi-group confirmatory factor analysis of the PROMIS Profile 29 provided evidence that measurement invariance largely holds between different languages. We observed significant differences in patient-reported health between countries, which could be partially explained by differences in overall ratings of health. The physical function and pain interference scales showed considerable floor effects in the normal population in all countries.

Conclusions: Scores derived from the PROMIS Profile 29 are largely comparable across the UK, France and Germany. Due to the use of plausible value imputation, the presented general population reference values can be compared to data collected with other PROMIS short forms or computer-adaptive tests.

Introduction

Assessing the patients' perspective about their health is vital when it comes to clinical decision making [1], in medical research [2] and in health policy making [3]. There are efforts to define standardized outcome measurement sets for certain diseases [4], but still one of the main barriers in the field of cross-national research in Europe is the availability of comparable, standardized and affordable health status measurements. Popular examples of instruments eliciting such measurements are the SF-36 [5, 6] and the EQ-5D [7]. However, currently only a handful of such instruments have been rigorously cross-culturally adapted and evaluated in their translated versions and fewer are freely available, hampering the ability to draw valid conclusions from results from multi-national research [8, 9].

In the last decade, the PROMIS initiative in the United States developed a large number of item banks for the assessment of diverse health domains including physical (e.g. Physical Functioning, Pain), mental (e.g. Depression, Anxiety), and social health (e.g. Role Functioning) [10–12]. These item banks were developed using Item Response Theory (IRT) methods which facilitate precise and efficient measurement across a wide range of each target domain [11, 13]. In order to be included in the final item banks, items underwent a rigorious evaluation process, including expert review, cognitive testing, and psychometric analysis to rule out measurement bias in regards to gender, age or disease status [14]. Importantly, the IRT methods used also allow established questionnaires to be anchored to the same scale, making data from different sources more easily comparable [15–18]. Also, IRT methods allow for a detailed examination of item properties, e.g. whether items which are translated into different languages perform equally well across translations [19], a process known as differential item functioning (DIF)[20].

Following the success of the PROMIS initiative an international effort to translate PROMIS instruments into various languages has been initiated to allow comparability of individual health statuses with a measurement that is independent of the country it is used in or the language it is translated into [9]. The translation process of a PROMIS measure follows a state-of-the-art approach, including forward and backward translation as well as an expert review, cognitive debriefing with patients to ensure content equivalence and final approval through the PROMIS Health Organization [14, 21]. Reports on some of the translations into French [22] and German [23–25] have been published elsewhere.

Some PROMIS item banks have been evaluated regarding language DIF, in particular Physical Functioning between US and Latino [26] as well as US and Dutch samples [27], Pain Interference between US and Dutch samples [28], Depression between US and German samples [29] and Social Health between US and Spanish-speaking samples [30]. While a substantial impact of language related DIF was only found in Physical Functioning between US and Latino [26], in others language-related DIF was negligible [27–29] or even absent [30], indicating that data collected on different language verisions of the measures can be compared directly and accurately.

PROMIS short forms are available for each domain [31, 32], and the PROMIS Profile 29 combines short forms of core domains in order to obtain a generic measure of health status, including aspects of physical, mental and social health. This instrument has the potential to fill an important gap in the toolkit for outcome measurement in cross-national studies and health systems performance frameworks [3].

The aims of this paper are to (1) investigate whether PROMIS Profile 29 scores can be compared between the UK, French and German versions, (2) evaluate if and to what extent differences in perceived health exist between UK, French, and German general populations and (3) finally to provide reference values from representative UK, French, and German general population samples in order to facilitate interpretation of PROMIS Profile domain scores in research and practice.

Methods

Samples

Data was collected in France (n = 1,501), Germany (n = 1,502) and the UK (n = 1,509) by an independent polling company (Ipsos) through their internet panel. There was no missing data as participants had to respond to each question in order to proceed with the survey.

Quota sampling was conducted to obtain general population samples from each country representative in regards to gender, age, occupation, region, and population density of the living place. To account for minor deviations from the marginal distributions, sample weights were calculated using the Random Iterative Method (RIM) to match the latest data available in each country (census 2011 for UK and Germany, census 2012 for France).

Measures

Sociodemographic information collected included gender, age, education, income, occupation, marital status, and household size. We obtained scores from the visual analog scale of the EQ-5D as a composite measure of overall self-rated health status [33].

The PROMIS Profile 29 is a generic patient reported outcome measure[34]. It combines short forms with four items each from seven PROMIS domains (depression, anxiety, physical function, pain interference, fatigue, sleep disturbance, and ability to participate in social roles and activities) and a single item on pain intensity, hence combining different aspects of health deemed appropriate for most adults [35]. The itembanks for each of the domains have been developed following state-of-the-art methods [36–41], a growing body of evidence shows their clinical usefulness [42–48]. Specifically, the PROMIS Profile 29 has been sufficiently sensitive to detect differences in quality of life across different samples of patients [49–54].

We used the language versions of the PROMIS Profile 29 for the UK, France and Germany as distributed by the PROMIS Health Organization. The translations of the respective underlying item banks have been conducted according to PROMIS scientific standards, including forward and backward translation, expert review, cognitive debriefing and final approval by PROMIS

US [14]. In the UK version one item “Are you able to do chores such as vacuuming or yard work?” had been previously adapted for use in the UK as “Are you able to do chores such as vacuuming housework or light gardening?” in the Stanford Health Assessment Questionnaire [55]. Besides that, no formal investigation of cultural acceptability of the US version in the UK was conducted.

Prior to data collection, we compared the wording of each item in the three language versions in a group of 5 health care professionals speaking at least two of the three languages fluently. We found no apparent differences in content in 20 items, slight differences in 6 items (EDANX01, EDANX40, EDANX41, EDDEP04, FATEXP41, PAININ22) and considerable differences in 3 items (PFA23, PFA53, EDANX53). For example item PF53 was judged to have a connotation of joy in its French translation (“Êtes-vous capable de faire des courses ou du shopping?”), which was absent in the English (“Are you able to run errands and shop?”) and German version (“Können Sie Ihre Besorgungen selbst machen und einkaufen gehen?”).

Statistical Analysis

Sociodemographics

Sociodemographic data are presented for each country. We tested for country differences of sociodemographic data using weighted chi-square tests for categorical variables and univariate analysis of variance for continuous variables.

Measurement Invariance

Assessment of measurement invariance is a prerequisite of valid latent variable comparisons across different languages. PROMIS measure development and DIF analysis has been done using unidimensional IRT models, but here we applied a multi-group confirmatory factor analysis (CFA) approach, as we intended to investigate measurement invariance of the PROMIS Profile 29 including the respective correlations between domains. Although this would be also possible using multidimensional IRT models, estimation of those is computational demanding and prone to numerical instability.

Seven domains of the PROMIS Profile 29 (Physical Function, Anxiety, Depression, Fatigue, Sleep Disturbance, Ability to participate in social roles and activities, Pain Interference) were modeled as correlated latent factors with the respective four items loading solely on that factor (see supplementary Figure 1). Unlike the other domains, pain intensity is assessed with a single indicator in the PROMIS Profile 29 and was therefore excluded from latent variable analysis. We used the weighted least squares estimator with mean and variance correction (WLSMV) given the ordinal response options.

At first, we fitted a configural invariant model to apply the same factorial structure to the data from each country. The configural invariant model was formulated using theta parameterization and identified by setting the means and variances of the latent factors to 0 and 1, respectively, item intercepts to 0 and residual variances to 1 [56]. Under theta parameterization the model has 4 relevant measurement parameters: thresholds, loadings, intercepts and residual variances [57]. Starting from the configural invariant model, we constrained parameters one at a time to

be the same over groups. Factor means and variances are comparable when threshold, loading and intercept invariance holds [56]. Identification constraints for each model were imposed following recent advice [56].

The overall fit of the models was assessed by chi-square statistics, the Comparative Fit Index (CFI, cut-off >.95) and the Root Mean Square Error of Approximation (RMSEA, cut-off <.08) [58, 59]. We used scaled chi-square difference tests [60, 61] to test whether introducing equality constraints decreased model fit. At large sample sizes, chi-square tests have an excellent power to detect small, possibly irrelevant effects [62], so it has been suggested to compare change in goodness-of-fit indices between those models. We report those changes, but proposed cutoffs are derived under limited simulation conditions and not consistent [63, 64]; it has been recommended to avoid their interpretation entirely when using WLSMV estimation [65].

Hence, we investigated the impact of potentially miss-specified parameter equality over groups on latent factor scores. Coming from a fully invariant model, we released threshold and loading equality constraints for each item at a time. Intercept and residual variance remained fixed at 0 and 1, respectively, across groups for identification. The 28 partial invariant models were tested against the fully invariant model using scaled chi-square difference tests and we report the change of latent factor estimates.

PROMIS Profile 29 scoring and plausible value imputation

We then converted raw sum scores of each domain of the PROMIS Profile 29 into standardized T-scores following the PROMIS scoring manual [35] and used these scores for further analysis. These scores are derived from IRT models of the respective outcomes calibrated in the US general population [31]. The underlying item parameters can be obtained through the PROMIS Assessment center (http://www.assessmentcenter.net). Although our analysis of measurement invariance does not test appropriateness of the US scoring algorithm, its application is highly relevant, because no country specific scoring algorithms exist so far and it is likely that US scoring will be used as a default.

Although frequently claimed as an advantage of IRT, the standard error of individual estimates of the latent trait is seldom taken into account in statistical analysis. We conducted multiple imputation of plausible values in order to account for the standard error of measurement of latent trait estimates and to obtain a continuous distribution of the latent variable for the whole sample [66–68]. We imputed 25 sets of data were each latent trait estimate was replaced by a random draw from a normal distribution N(score estimate, standard error of measurement), ran all analysis in each imputed dataset and pooled estimates subsequently according to Rubin's rule [69]. Imputation of plausible values cannot increase precision of individual estimates, but results in more appropriate estimation on the sample level [68].

Differences in perceived health

In order to investigate differences in perceived health between the three countries, we fitted linear regression models for each of the 7 domains separately and used them to estimate crude mean scores and the respective confidence intervals. We then expanded those models by including covariates in order to explain the observed health differences between countries. At

first, we included the available sociodemographic variables (age, gender, education, income, occupation, marital status, household size) and in a second step, we included perceived overall health (EQ-5D VAS). Sampling weights were taken into account.

General population reference values

Linear regression models the mean of the variable of interest – quantile regression extends this approach to model arbitrary quantiles [70]. Therefore, it allows estimation of the value of an outcome at any quantile (for example the Physical Function score that 90% of the sample achieve). This value comes with an estimate of uncertainty (standard error) and one can investigate the impact of a given a set of predictors on that value.

We used quantile regression [70] to model the 10^{th} to 90^{th} crude domain specific percentile and their respective standard errors in each of the imputed datasets. Sampling weights were taken into account. Standard errors were calculated using an asymptotic approximation accounting for non-iid errors [70]. The estimated percentiles and their respective standard errors were pooled over the imputed datasets according to Rubin's rule. Those pooled estimates were then used to calculate 95% confidence intervals assuming a normal distribution.

We fitted a second set of quantile regression models including age, gender, income, and education variables to investigate the influence of these variables on the distribution of the outcome. Furthermore, these models allow predicition of adjusted percentiles which can be used as sample specific reference values.

For Pain Intensity, we report unadjusted means and standard deviations along with cumulative percentages for each country.

Mplus 7.4 was used to estimate weighted confirmatory analysis for ordinal data. All other analysis were conducted in the R Statistical Programming Environment [71], using packages 'weights' for calculation of weighted chi-square tests [72], 'quantreg' [73] for quantile regression and 'Amelia' [74] to combine results from multiple imputed datasets.

Results

Sociodemographics

Table 1 shows the sociodemographic data from the different samples. While the gender ratio is similar in each country, there are, as expected since sample size is large, statistically significant differences in age, education, occupation, income, household size and marital status. Also, the mean of self-reported health differs between countries by about a quarter standard deviation (5 points on the scale ranging from 0 to 100) and is generally considerably (up to 10 points) lower than in earlier reports of nationally representative data [75].

*** Insert Table 1 ***

Measurement Invariance

The configural model did not fit the observed data exactly as indicated by the significant chi²-test, whereas CFI (>.95) and RMSEA (<.08) show acceptable approximate fit of the hypothesized factor structure. Scaled chi² difference tests indicate that constraining thresholds, loadings, intercepts and residual variances lead to significantly worse model fit, whereas fit indices decrease only slightly. Assuming intercept invariance results in the largest chi² difference (see Table 2).

*** Insert Table 2 ***

Releasing item's threshold and loading parameters resulted in significantly better fit of the respective model in 27 of 28 cases (see Table 3). In 3 (France) respectively 6 (Germany) models releasing item parameters lead to a change in the respective domain factor score larger than 0.05 and in 1 case (PFA23), factor scores changed by more than 0.10. Given that the factor score estimates come with a standard error of about 0.05, these changes appear to be reasonably small in most cases. It is also worth noting that in all domains the changes introduced by releasing items seem to cancel each other out.

*** Insert Table 3 ***

Interestingly, the items identified to have differences in their wording across languages do not stand out in this analysis and we were not able to identify any difference in the wording of item PFA23 ("Are you able to go up and down stairs at a normal pace?", "Können Sie mit normaler Geschwindigkeit Treppen hoch- und runtergehen?", "Êtes-vous capable de monter et descendre les escaliers à un rythme normal?"), for which reestimation of item parameters results in the largest change of the factor score.

Differences in perceived health

Figure 1 shows the crude and adjusted mean scores and their respective confidence intervals. Compared to the US calibration sample mean of 50 and a standard deviation of 10, general population data from the UK, Germany, and France differed by up to 4 points. Specifically, the French sample reported lower scores for Depression and Fatigue and a better Ability to participate in social roles and activities compared to the UK and German samples. Given the T-score metric with m = 50, sd = 10 in the US general population, this translates to small to medium effect sizes. Adjusting for sociodemographic variables did not change that pattern considerably, but accounting for sample differences in health status (EQ-5D VAS) decreased in particular the large mean differences in Depression and Fatigue between countries. This suggests that observed differences are associated with actual differences in perceived health between samples.

*** Insert Figure 1 ***

In most of the domains, we observed considerable floor effects (Anxiety (27.6%), Depression (38.2%), Fatigue (20.6%)) and ceiling effects for functioning (Physical Function (67.3%), Ability to participate in social roles and activities (27.8%), and Pain interference (47.6%)), indicating that these short forms ability to differentiate between healthy persons is rather low. For Sleep

Disturbance we found no pronounced floor effect with only 4.8% of participants responding with the lowest possible score.

General population reference values

The domain and country specific T-scores for each percentile are presented Table 4. For example, an observed raw score of 10 on the anxiety scale corresponds to a T-value of 59.5 [35]. Since the 70th percentile is 58.9 (UK), 58.1 (France) and 57.4 (Germany), respectively, it can be followed that 70% of the general population achieve lower anxiety levels. Given the confidence interval of such an individual estimate (59.5+/-2*2.6 = 54.4 to 64.6)[35], it is likely that this person's anxiety exceeds 50% of the population's anxiety, but is less than the anxiety of the highest 10 percent (90th percentile).

*** Insert Table 4 ***

Figure 2 shows the quantile regression estimates for the the 7 PROMIS Profile domains from the adjusted models with their respective confidence intervals. Percentile estimates of Anxiety, Depression and Fatigue scores are elevated in women compared to men. Higher age is associated with lower scores in Physical Function and higher scores for Pain Interference but also with lower scores for Anxiety, Depression and Fatigue. While there appears to be no difference between low and medium levels of education, high level of education seems to be associated with less symptoms and higher performance (e.g. the 10% percentile with high education scores about 3 points higher in Physical Function compared to low education). Income is associated with better Physical Function, less Pain Interference, less Anxiety, Depression and Fatigue and increased Ability to participate in social roles. Notably is that the country difference in Ability to participate in social roles and activities is particularily due rather high scores of the lower percentiles in French and and low scores of higher percentiles in German populations. Anxiety and Depression seem to follow a narrower distribution in the German sample.

*** Insert Figure 2 ***

The quantile regression models make it possible to calculate reference values for specific subsets of the general population. For convenience, we offer a web-application to obtain percentiles given country, age, gender, education and income (http://www.common-metrics.org/PROMIS_Profile_29_General_Population.php).

We found no significant mean difference in self-reported pain intensity between countries (UK: 2.61 [2.49; 2.74], France: 2.57 [2.44; 2.70], Germany: 2.62 [2.49; 2.75]). We present the cumulative percentages of the raw scores and the respective confidence intervals in each country in Table 5.

*** Insert Table 5 ***

Discussion

Using general population data we investigated measurement invariance of the PROMIS Profile 29 and differences in self reported health in samples from UK, France and Germany.

Furthermore, we present reference values for the 7 PROMIS Profile domains of depression, anxiety, physical function, pain interference, fatigue, sleep disturbance, and ability to participate in social roles and activities from general population in the UK, France, and Germany.

Multi-group confirmatory factor analysis showed that the hypothesized model structure with 7 correlated health domains fits the data reasonably well. Imposing equality constraints on thresholds, loadings, intercepts and residual variances across groups resulted in significant worse fit. Goodness-of-fit measures seem excellent, but those shall not be used to investigate measurement invariance under WLSMV estimation [65]. Analysis of the impact of releasing thresholds and loadings for one item at a time showed that only for one item latent factor scores were considerably influenced. This lets us conclude that the three versions of the PROMIS Profile 29 have largely comparable measurement properties. When investigating small effects between countries it might be advisable to take into account the possibility of some measurement bias.

We found some differences in self-reported health between UK, France and Germany. Our analysis revealed that these could be not explained by differences in sociodemographic variables between samples, but at least partially by differences in global ratings of health. The effects from sociodemographic variables are not surprising – we observed worse function and more symptoms in the older and less educated. It is important to note that the adjustment for the EQ-5D visual analogue scale could mask real health differences between the countries – however, from a psychometric perspective it can be reasoned that the observed differences in PROMIS Profile 29 scores does indeed have more to do with real health differences across the samples as with differences in the different translations of the PROMIS Profile 29. However, a potential risk of bias is that measurement equivalence of the EQ-5D VAS has not been established.

Depending on scale orientation, we observed strong floor respectively ceiling effects on most scales, indicating that healthy persons can be hardly differentiated with the PROMIS Profile 29. This has also been recently reported to be true in patient samples as well [76]. In particular the Physical Function and Pain Interference Short Forms seem to be insufficiently tailored to match the true distribution of the respective latent variables in the general population. In general, floor and ceiling effects can cause a serious lack of responsiveness. For future studies we would therefore advise the use of short forms containing more ‘difficult’ items (i.e. items which require exceptionally high levels of functional ability to affirm) in both domains. One of the advantages of the IRT-based PROMIS framework is that such a tailoring can be done in a straightforward fashion and that resulting scores would be still comparable.

Percentiles based on plausible value imputation allow straightforward, reference based interpretation of scores for each of the 7 PROMIS Profile domains. A major advantage of the PROMIS scales is that they are anchored to a meaningful numerical value – 50 is the mean score of the US general population and close estimates were observed for 50 percentiles all domains across the three countries. However, given that each domain in the PROMIS Profile 29 is only measured with 4 items, estimates on an individual patient level come with large standard errors. Differences in percentiles across countries appear to be small compared with the uncertainty of the latent trait estimates. If the individual level is of interest, one should therefore make use of

more precise PROMIS measures, such as longer short forms or CATs. Since the distribution of plausible values is a consistent estimator of the true latent variable distribution [77] the reported percentiles reflect the latent variable distribution. Hence, one can use these not only as reference values for latent trait estimates derived with the PROMIS Profile 29, but also from estimates of the latent trait derived from other short forms or computer-adaptive tests of the respective domains. Nonetheless, common criteria for CATs are SE of 3.2 or 2.2 (given the T-metric with mean of 50 and a standard deviation of 10, as used in the PROMIS framework), resembling a reliability of .90 and .95. These translate into confidence intervals of individual estimates of +/- 6.3 and 4.3, which still exceed differences in percentiles across countries.

Strengths and Limitations

General population data was collected using the same methods in all countries and quota sampling was used to obtain representative samples. Unlike truly random sampling plans, quota sampling might result in biased samples, since only marginal distributions of certain variables are like those from the general population – other variables might be differently distributed. Furthermore, there might be a selection bias incurred by the nature of online polling. This could be the reason why participants in this study rate their health in the EQ-5D VAS lower than expected. Nonetheless, quota sampling has also been used for the PROMIS calibration sample [78].

A major strength of this paper is the use of plausible value imputation. This approach allows using the presented general population percentiles not only for estimates derived with the PROMIS Profile 29, but also for data obtained with other PROMIS Short Forms or CATs for the respective domains. Plausible value imputation has been shown to consistently estimate the true latent variable distribution [77] and results in less biased effect estimates compared to sum scores, e.g. in analysis of RCT data [79]. Although the underlying IRT model for plausible value imputation was estimated in US samples, current evidence suggests that differential item functioning does not interfere with cross-national comparisons using the different PROMIS measures [27–30]. A further possible limitation of the plausible value approach as implemented is that we assumed a normal distribution of the true scores of latent variable estimates instead of using the actual likelihood and that sum score IRT estimate were used instead of individual response patterns. However, differences between estimates from specific response patterns and sum scores have been reported to be small [80].

A crucial limitation of this study is that we used the US scoring algorithm, although it is unclear whether it is appropriate over all domains and languages. An investigation of the appropriateness of this specific measurement model would require the collection of the full item banks and US samples, which was beyond the scope of our project. Furthermore, given the number of item banks developed in PROMIS and the sample size that would be necessary to calibrate the respective IRT models in each language it seems unlikely that other language specific scoring algorithms will be available anytime soon. We found in an earlier study that using parameters from an existing IRT model in other samples yields comparable results to reestimated measurement models [81]. This makes us somewhat confident that an European scoring algorithm would be the similar to the US scoring algorithm. It should be kept in mind

that although we were able to provide evidence that scores between UK, French and German samples are comparable, this might not hold for comparisons with US data.

A further limitation of our study is that we relied on multigroup confirmatory factor analysis models alone to investigate measurement invariance. While those models seem straightforward under maximum likelihood estimation, they impose unexpected difficulties in the case ordinal response data. Only recently it has been shown that identification constraints must be carefully adapted when introducing equality constraints [56], implying misspecifications of models in earlier studies. Furthermore, it has been advised against common practice of interpreting change of goodness-of-fit [65]. However, assessment of DIF in an IRT framework would lead into the same problems of extremely powerful chi²-tests that detect irrelevant DIF and the reliance on approximate goodness-of-fit measures, where a threshold for clinical relevance is hard to define.

Conclusion

Our analysis reveals that the PROMIS Profile 29 is a suitable generic instrument to measure health status in cross-national studies between the UK, France and Germany. Interpretation of PROMIS Profile 29 scores for researchers and clinicians based on population percentiles is straightforward and since we modeled the underlying latent variable distributions using a plausible value approach percentiles can also serve as preliminary general population reference for other PROMIS short forms and CATs as well.

The mean and standard deviation of all PROMIS scales are anchored on the US population; this could be misleading when interpreting scores from other countries as the general population could have considerably lower or higher scores. On the other hand, one could argue those mean differences appear to be small. An open question for the future therefore remains: should we anchor scales based on the US general population, the respective country's population or even on a global level?

Compliance with Ethical Standards

Funding: This study was funded by the Centre Virchow-Villerme.

Conflict of Interest: Authors declare that they have no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

1. Basch, E. (2017). Patient-Reported Outcomes — Harnessing Patients' Voices to Improve Clinical Care. *New England Journal of Medicine*, *376*(2), 105–108. doi:10.1056/NEJMp1002530
2. Snyder, C. F., Jensen, R. E., Segal, J. B., & Wu, A. W. (2013). Patient-reported Outcomes (PROs) Outcomes Research. *Medical Care*, *51*(8 Suppl 3), 73–79.
3. Black, N., Burke, L., Forrest, C. B., Ravens Sieberer, U. H., Ahmed, S., Valderas, J. M., … Alonso, J. (2016). Patient-reported outcomes: pathways to better health, better services, and better societies. *Quality of Life Research*, *25*(5), 1103–1112. doi:10.1007/s11136-015-1168-3
4. McNamara, R. L., Spatz, E. S., Kelley, T. A., Stowell, C. J., Beltrame, J., Heidenreich, P., … Lewin, J. (2015). Standardized Outcome Measurement for Patients With Coronary Artery Disease: Consensus From the International Consortium for Health Outcomes Measurement (ICHOM). *Journal of the American Heart Association*, *4*(5), e001767-. doi:10.1161/JAHA.115.001767
5. Ware, J. E., Kosinski, M., Gandek, B., & Aaronson, N. (1998). The Factor Structure of the SF-36 Health Survey in 10 Countries : Results from the IQOLA Project. *Journal of clinical epidemiology*, *51*(11), 1159–1165.
6. Bullinger, M., Alonso, J., Apolone, G., Leplège, A., & Sullivan, M. (1998). Translating Health Status Questionnaires and Evaluating Their Quality : The IQOLA Project Approach. *Journal of clinical epidemiology*, *51*(11), 913–923.
7. Szende, A., Janssen, B., & Cabases, J. (2014). *Self-Reported Population Health: An InternationalPerspective based on EQ-5D*. Dordrecht: Springer. doi:10.1007/978-94-007-7596-1
8. 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), 3186–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11124735
9. Alonso, J., Bartlett, S. J., Rose, M., Aaronson, N., 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
10. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K. F., Reeve, B. B., … Rose, M. (2007). Developing the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*(5), 3–11. Retrieved from http://journals.lww.com/lww-medicalcare/Abstract/2007/05001/Developing_the_Patient_Reported_Outcomes.1.aspx
11. Cella, D., Yount, S., Rothrock, N., & Gershon, R. (2007). 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. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2829758/
12. 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), S22-31. doi:10.1097/01.mlr.0000250483.85507.04
13. Böhnke, J. R., & Lutz, W. (2014). Using item and test information to optimize targeted assessments of psychological distress. *Assessment*, *21*(6), 679–93. doi:10.1177/1073191114529152
14. Patient-Reported Outcomes Measurement Information System. (2013). PROMIS Instrument Development and Validation Scientific Standards Version 2.0. Retrieved

March 20, 2016, from http://www.nihpromis.org/Documents/PROMISStandards_Vers2.0_Final.pdf

15. 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), 513–27. doi:10.1037/a0035768
16. Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders*, *28*(1), 88–96. doi:10.1016/j.janxdis.2013.11.006
17. Fischer, H. F., Tritt, K., Klapp, B. F., & Fliege, H. (2011). How to compare scores from different depression scales: equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using Item Response. *International Journal of Methods in Psychiatric Research*, *20*(4), 203 – 214. doi:10.1002/mpr
18. 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
19. Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.
20. Holland, P., & Wainer, H. (2012). *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.
21. Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., … Butt, Z. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research*. doi:10.1007/s11136-012-0344-y
22. Bartlett, S. J., Witter, J., Cella, D., & Ahmed, S. (2017). Montreal accord on patient-reported outcomes use series-paper 6: Creating national initiatives to support development and use-the PROMIS example. *Journal of Clinical Epidemiology*. doi:10.1016/j.jclinepi.2017.04.015
23. 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.
24. Farin, E., Nagl, M., Gramm, L., Heyduck, K., & Glattacker, M. (2013). Development and evaluation of the PI-G: a three-scale measure based on the German translation of the PROMIS pain interference item bank. *Quality of Life Research*. doi:10.1007/s11136-013-0575-6
25. Liegl, G., Rose, M., Correia, H., Fischer, H. F., Kanlidere, S., Mierke, 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. *Clinical Rehabilitation*, 26921551771429. doi:10.1177/0269215517714297
26. 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), 1819–30. doi:10.1007/s11136-012-0292-6
27. Oude Voshaar, M. A. H., ten Klooster, P. M., Glas, C., 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
28. 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
29. Fischer, H. F., 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. *International Journal of Methods in Psychiatric Research*, (epub first).
30. 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), 490–9. doi:10.1037/hea0000055
31. 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), 125–36. doi:10.1007/s11136-009-9560-5
32. Cella, D., Gershon, R., Lai, J.-S., Choi, S. W., Yount, S., Rothrock, N., … Rose, M. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16 Suppl 1*(5 Suppl 1), 133–41. doi:10.1007/s11136-007-9204-6
33. van Reenen, M., & Janssen, B. (2015). EQ-5D-5L User Guide - Basic information on how to use the EQ-5D-5L instrument. Retrieved from http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-5L_UserGuide_2015.pdf
34. Valderas, J. M., & Alonso, J. (2008). Patient reported outcome measures: a model-based classification system for research and clinical practice. *Quality of Life Research*, *17*(9), 1125–35. doi:10.1007/s11136-008-9396-4
35. Patient-Reported Outcomes Measurement Information System. (2013). PROMIS Short Form Scoring Manual. Retrieved March 21, 2016, from http://www.assessmentcenter.net/documents/PROMIS Profile Scoring Manual.pdf
36. Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W., Choi, S., Revicki, D. A., … Callahan, L. (2010). Development of A Promis Item Bank to Measure Pain Interference. *Pain*, *150*(1), 173–182. doi:10.1016/j.pain.2010.04.025.Development
37. Hahn, E. A., Devellis, R. F., Bode, R. K., Garcia, S. F., Castel, L. D., Eisen, S. V, … Cella, D. (2010). Measuring social health in the patient-reported outcomes measurement information system (PROMIS): item bank development and testing. *Quality of Life Research*, *19*(7), 1035–44. doi:10.1007/s11136-010-9654-0
38. Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, *92*(10 SUPPL.), S20–S27. doi:10.1016/j.apmr.2010.08.033
39. 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), 263–83. doi:10.1177/1073191111411667
40. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of clinical epidemiology*, *67*(5), 516–26. doi:10.1016/j.jclinepi.2013.10.024
41. Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., … Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep*, *33*(6), 781–92. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2880437&tool=pmcentrez&rendertype=abstract

42. Schalet, B. D., Hays, R. D., Jensen, S. E., Beaumont, J. L., Fries, J. F., & Cella, D. (2016). Validity of PROMIS physical function measures in diverse clinical samples. *Journal of Clinical Epidemiology*, *73*, 112–118. doi:10.1016/j.jclinepi.2015.08.039
43. Schalet, B. D., Pilkonis, P. A., Yu, L., Dodds, N., Johnston, K. L., Yount, S., … Cella, D. (2016). Clinical validity of PROMIS Depression, Anxiety, and Anger across diverse clinical samples. *Journal of Clinical Epidemiology*, *73*, 119–127. doi:10.1016/j.jclinepi.2015.08.036
44. Cella, D., Lai, J.-S., Jensen, S. E., Christodoulou, C., Junghaenel, D. U., Reeve, B. B., & Stone, A. A. (2016). Clinical Validity of the PROMIS® Fatigue Item Bank across Diverse Clinical Samples. *Journal of Clinical Epidemiology*, *73*(2016), 128–134. doi:10.1016/j.jclinepi.2015.08.037
45. Cook, K. F., Jensen, S. E., Schalet, B. D., Beaumont, J. L., Amtmann, D., Czajkowski, S., … Cella, D. (2016). PROMIS® Measures of Pain, Fatigue, Negative Affect, Physical Function and Social Function Demonstrate Clinical Validity across a Range of Chronic Conditions. *Journal of Clinical Epidemiology*, *73*, 89–102. doi:10.1016/j.jclinepi.2015.08.038
46. Hahn, E. A., Beaumont, J. L., Pilkonis, P. A., Garcia, S. F., Magasi, S., DeWalt, D. A., & Cella, D. (2016). The PROMIS satisfaction with social participation measures demonstrate responsiveness in diverse clinical populations. *Journal of Clinical Epidemiology*, *73*, 135–141. doi:10.1016/j.jclinepi.2015.08.034
47. Stone, A. A., Broderick, J. E., Junghaenel, D. U., Schneider, S., & Schwartz, J. E. (2015). PROMIS fatigue, pain intensity, pain interference, pain behavior, physical function, depression, anxiety, and anger scales demonstrate ecological validity. *Journal of Clinical Epidemiology*, *74*, 194–206. doi:10.1016/j.jclinepi.2015.08.029
48. Askew, R. L., Cook, K. F., Revicki, D. A., Cella, D., & Amtmann, D. (2016). Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. *Journal of Clinical Epidemiology*, *73*, 103–111. doi:10.1016/j.jclinepi.2015.08.035
49. Beaumont, J. L., Cella, D., Phan, a T., Choi, S., Liu, Z., & Yao, J. C. (2012). Comparison of health-related quality of life in patients with neuroendocrine tumors with quality of life in the general US population. *Pancreas*, *41*(3), 461–466. doi:10.1097/MPA.0b013e3182328045
50. Craig, B. M., Reeve, B. B., Brown, P. M., Cella, D., Hays, R. D., Lipscomb, J., … Revicki, D. A. (2014). US valuation of health outcomes measured using the PROMIS-29. *Value in Health*, *17*(8), 846–853. doi:10.1016/j.jval.2014.09.005
51. Pearman, T. P., Beaumont, J. L., Cella, D., Neary, M. P., & Yao, J. (2016). Health-related quality of life in patients with neuroendocrine tumors: an investigation of treatment type, disease status, and symptom burden. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*. doi:10.1007/s00520-016-3189-z
52. Yount, S. E., Beaumont, J. L., Chen, S.-Y., Kaiser, K., Wortman, K., Van Brunt, D. L., … Cella, D. (2016). Health-Related Quality of Life in Patients with Idiopathic Pulmonary Fibrosis. *Lung*, *194*(2), 227–234. doi:10.1007/s00408-016-9850-y
53. Hinchcliff, M., Beaumont, J. L., Thavarajah, K., Varga, J., Chung, A., Podlusky, S., … Cella, D. (2011). Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item Health Profile and Functional Assessment of Chronic Illness Therapy-Dyspnea short form. *Arthritis care & research*, *63*(11), 1620–8. doi:10.1002/acr.20591
54. Hinchcliff, M., Beaumont, J. L., Carns, M., Podlusky, S., Thavarajah, K., Varga, J., … Chang, R. W. (2015). Longitudinal evaluation of PROMIS-29 and FACIT-Dyspnea short forms in systemic sclerosis. *Journal of Rheumatology*, *42*(1), 64–72. doi:10.1530/ERC-14-

0411.Persistent
55. Kirwan, J. R., & Reeback, J. S. (1986). Using a modified Stanford Health Assessment Questionnaire to access disability in UK patients with rheumatoid arthritis. *British Journal of Rheumatology*, *25*, 206–209. doi:10.1093/rheumatology/25.2.206
56. Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, *81*(4), 1014–1045. doi:10.1007/s11336-016-9506-0
57. Millsap, R. E., & Tein, J. Y. (2004). Multivariate Behavioral Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903
58. Brown, T. A., & Kenny, D. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.
59. 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
60. Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514.
61. Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2).
62. Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*(1), 55–75. doi:10.1146/annurev-soc-071913-043137
63. Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. doi:10.1080/10705510701301834
64. Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. doi:10.1037/0021-9010.93.3.568
65. Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167–180. doi:10.1080/10705511.2014.882658
66. Glas, C., Geerlings, H., van de Laar, M. A. F. J., & Taal, E. (2009). Analysis of longitudinal randomized clinical trials using item response models. *Contemporary clinical trials*, *30*(2), 158–70. doi:10.1016/j.cct.2008.12.003
67. Gorter, 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 Methodology2*, 1–12. doi:10.1186/s12874-015-0050-x
68. Levy, R., & Mislevy, R. J. (2016). *Bayesian Psychometric Modeling*. Boca Raton: CRC Press.
69. Marshall, A., Altman, D. G., Holder, R. L., & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC medical research methodology*, *9*, 57. doi:10.1186/1471-2288-9-57
70. Hao, L., & Naiman, D. Q. (2007). *Quantile Regression*. Thousand Oaks: Sage Publications.
71. R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
72. Pasek, J. (2016). weights: Weighting and Weighted Statistics. *R package*. Retrieved from http://cran.r-project.org/package=weights
73. Koenker, R. (2016). quantreg: Quantile Regression. *R package*. Retrieved from http://cran.r-project.org/package=quantreg
74. Honaker, J., King, G., & Blackwell, M. (2011). AMELIA II : A Program for Missing Data.

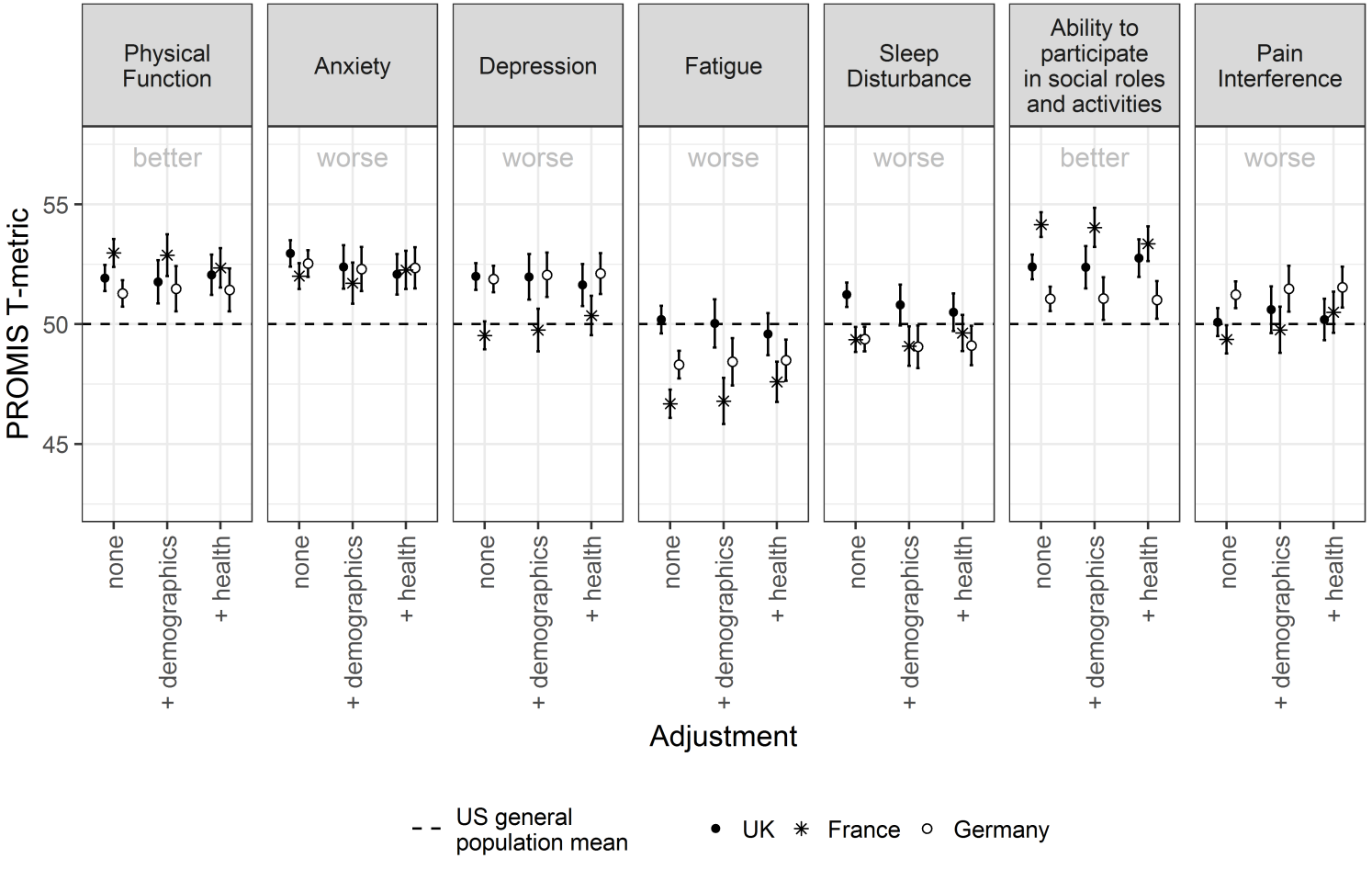
Journal Of Statistical Software, *45*(7), 1–54. doi:10.1.1.149.9611

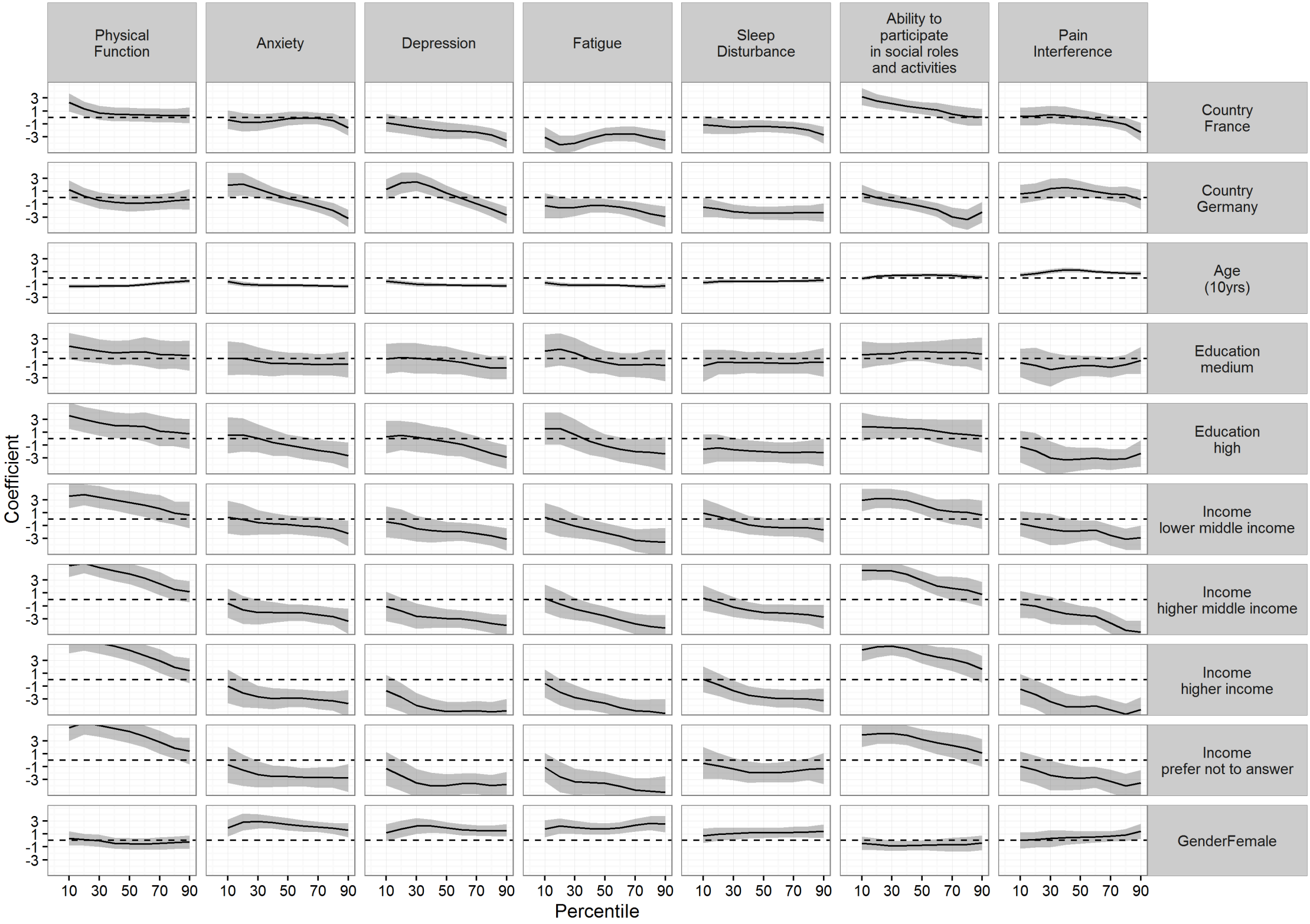
75. Janssen, B., & Szende, A. (2014). Population Norms for the EQ-5D. In A. Szende, B. Janssen, & J. Cabases (Eds.), *Self-Reported Population Health: An International Perspective based on EQ-5D* (pp. 19–30). Dordrecht: Springer. doi:10.1007/978-94-007-7596-1
76. Katz, P., Pedro, S., & Michaud, K. (2016). Performance of the PROMIS 29-Item Profile in Rheumatoid Arthritis, Osteoarthritis, Fibromyalgia, and Systemic Lupus Erythematosus. *Arthritis Care & Research*, *epub*. doi:10.1002/acr.
77. Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from Plausible Values? *Psychometrika*, (Ml). doi:10.1007/s11336-016-9497-x
78. Cella, D., Riley, W., Stone, A., Northrock, N., Reeve, B. B., Yount, S., … Hays, R. D. (2010). 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
79. Gorter, R., Fox, J.-P., Apeldoorn, A., & Twisk, J. (2016). The influence of measurement model choice for randomized controlled trial results. *Journal of Clinical Epidemiology*. doi:10.1016/j.jclinepi.2016.06.011
80. Thissen, D., & Wainer, H. (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.
81. Liegl, G., Wahl, I., Berghöfer, A., Nolte, S., Pieh, C., Rose, M., & Fischer, H. F. (2016). Using PHQ-9 item parameters of a common metric resulted in similar depression scores compared to independent IRT model reestimation. *Journal of Clinical Epidemiology*, *71*, 25–34. doi:10.1016/j.jclinepi.2015.10.006

Figure Legends

Figure 1: Estimated mean scores with 95% confidence interval from linear regression analysis. Unadjusted model includes country alone, adjustment for demographic includes age, gender, education, income, occupation, marital status and household size as covariates; adjustment for health ratings additionally ratings of the EQ-5D visual analogue scale

Figure 2: Effect estimates and respective 95% confidence intervals for the impact of country, age, gender, education and income on T-scores of the 10^{th} to 90^{th} percentiles. For example, the 10^{th} percentile of the French general population scores about 2.5 points higher in Physical Function then the UK general population, while the 30^{th} to 90^{th} percentile score almost identically (upper left panel)


Physical Function
Anxiety
Depression
Fatigue
Sleep Disturbance
Ability to participate in social roles and activities
Pain Interference
better
worse
worse
worse
worse
better
worse
PROMIS T-metric
55
50
45
none
+ demographics
+ health
Adjustment
US general population mean
UK
France
Germany


Physical Function
Anxiety
Depression
Fatigue
Sleep Disturbance
Ability to participate in social roles and activities
Pain Interference
Country France
Country Germany
Age (10yrs)
Education medium
Education high
Income lower middle income
Income higher middle income
Income higher income
Income prefer not to answer
GenderFemale
Coefficient
3
1
-1
-3
10
30
50
70
90
Percentile

		UK	France	Germany	p
Gender	Male	734.6 (48.7%)	714.5 (47.6%)	728.9 (48.5%)	
	Female	774.4 (51.3%)	786.5 (52.4%)	773.1 (51.5%)	.814
Age groups	18-29	306.3 (20.3%)	276.2 (18.4%)	250.7 (16.7%)	
	30-39	247.6 (16.4%)	249.2 (16.6%)	215.7 (14.4%)	
	40-49	271.3 (18.0%)	270.2 (18.0%)	274.9 (18.3%)	
	50-59	243.8 (16.2%)	256.7 (17.1%)	275.3 (18.3%)	
	60-69	209.8 (13.9%)	213.1 (14.2%)	199.3 (13.3%)	
	70 +	230.1 (15.3%)	235.7 (15.7%)	286.1 (19.0%)	.045
Education[1]	Low	122.3 (8.1%)	113.0 (7.5%)	101.2 (6.7%)	
	Medium	734.4 (48.7%)	640.5 (42.7%)	792.5 (52.8%)	
	High	652.3 (43.2%)	747.5 (49.8%)	608.2 (40.5%)	<0.001
Occupation	Managers and Professionals	317.5 (21.0%)	184.0 (12.7%)	186.4 (12.4%)	
	Technicians, Clerks, Service workers	374.7 (24.8%)	350.3 (24.1%)	437.2 (29.1%)	
	Workers, Elementary occupations, Armed forces	217.7 (14.4%)	241.0 (16.6%)	256.4 (17.1%)	
	Inactive/Unemployed	599.0 (39.7%)	675.7 (46.6%)	622.0 (41.4%)	<.001
Income[2]	Lower income	253.9 (16.8%)	188.2 (12.5%)	289.1 (19.3%)	
	Lower middle income	305.6 (20.3%)	288.4 (19.2%)	270.8 (18.0%)	
	Higher middle income	462.2 (30.6%)	412.6 (27.5%)	325.9 (21.7%)	
	Higher income	315.6 (20.9%)	368.9 (24.6%)	362.6 (24.1%)	
	Prefer not to answer	171.7 (11.4%)	242.8 (16.2%)	253.5 (16.9%)	<.001
Householdsize	mean (sd)	2.49 (1.15)	2.53 (1.14)	2.13 (1.05)	<.001
Marital Status	Never Married (Single)	394.0 (26.1%)	304.4 (20.3%)	371.5 (24.7%)	
	Domestic Partner (Living as a couple)	207.7 (13.8%)	226.5 (15.1%)	145.1 (9.7%)	
	Married / Civil Partnership	706.3 (46.8%)	798.4 (53.2%)	718.5 (47.8%)	
	Separated	27.7 (1.8%)	29.3 (1.9%)	22.1 (1.5%)	
	Divorced	116.1 (7.7%)	102.3 (6.8%)	168.3 (11.2%)	
	Widowed	57.3 (3.8%)	40.1 (2.7%)	76.5 (5.1%)	<.001
EQ-5D VAS	mean (sd)	71.65 (21.22)	76.50 (19.56)	73.24 (21.72)	<.001

[1]: Low: formal education up to the age of 15 (UK: Secondary school without General National Vocational Qualification, France: Collège (BREVET), Germany: qualifizierender Hauptschulabschluss), Medium: formal education up to the age of 18/19 (UK: GCE Advanced Level, France: Baccalauréat, Germany: abeschlossene Lehre/Fachabitur), High: further formal education (UK: Bachelor's/Master's Degree, France: Baccalauréat + 2 to 5yrs, Germany: Abitur, Diplom)
[2]: Yearly net household income: Lower income (UK: up to 15,000£, France/Germany: up to 15000€), Lower middle income (UK: 15000£ to 24999£, France/Germany: 15001€ to 24000€), Higher middle income (UK: 25000£ to 44.999£, France/Germany: 24001€ to 36000€), Higher income (UK: 45000£ and over, France/Germany: 36001€ and above)

Table 1: Sociodemographic data of the weighted general population samples

	Level of invariance	Chi²	df	p-value	Scaled Chi² Difference Test ΔChi²	Δdf	p-value	CFI	RMSEA [95% CI]
WLSMV	Configural	4585.5	987	<.001				0.994	0.049 [0.048 - 0.051]
	Constrained thresholds	4832.7	1099	<.001	321.3	112	<.001	0.994	0.048 [0.046 - 0.049]
	Constrained thresholds & loadings	4966.2	1141	<.001	231.8	42	<.001	0.993	0.047 [0.046 - 0.049]
	Constrained thresholds, loadings & intercepts	5723.7	1183	<.001	941.8	42	<.001	0.992	0.051 [0.049 - 0.052]
	Constrained thresholds, loadings, intercepts & residual variances (full invariance)	5226.3	1239	<.001	239.3	56	<.001	0.993	0.046 [0.045 - 0.048]

Table 2: Chi² statistics and goodness of fit of multigroup confirmatory factor analysis models with different levels of measurement invariance

		Scaled Chi² Difference Test			Latent factor means France			Latent factor means Germany		
Domain	**Item**	**ΔChi²**	**Δdf**	**p-value**	**invariant model**	**released model**	**Δ**	**invariant model**	**released model**	**Δ**
Physical Function	PFA11	33.501	10	<0.001	-0.015	-0.008	0.007	-0.292	-0.332	-0.040
	PFA21	66.150	10	<0.001		0.116	0.131		-0.158	0.134
	PFA23	26.341	10	0.003		-0.073	-0.058		-0.317	-0.025
	PFA53	61.581	10	<0.001		-0.050	-0.035		-0.354	-0.062
Anxiety	EDANX01	84.451	10	<0.001	-0.018	-0.036	-0.018	0.012	-0.046	-0.058
	EDANX40	165.255	10	<0.001		-0.047	-0.029		0.049	0.037
	EDANX41	28.915	10	0.001		-0.032	-0.014		0.016	0.004
	EDANX53	85.827	10	<0.001		0.037	0.055		0.016	0.004
Depression	EDDEP04	59.150	10	<0.001	-0.153	-0.199	-0.046	0.057	0.066	0.009
	EDDEP06	12.940	10	0.227		-0.151	0.002		0.069	0.012
	EDDEP29	95.997	10	<0.001		-0.124	0.029		0.028	-0.029
	EDDEP41	38.499	10	<0.001		-0.144	0.009		0.060	0.003
Fatigue	HI7	84.204	10	<0.001	-0.346	-0.298	0.048	-0.174	-0.114	0.060
	AN3	40.896	10	<0.001		-0.361	-0.015		-0.198	-0.024
	FATEXP41	75.329	10	<0.001		-0.381	-0.035		-0.191	-0.017
	FATEXP40	84.992	10	<0.001		-0.318	0.028		-0.177	-0.003
Sleep Disturbance	Sleep109	182.114	10	<0.001	-0.220	-0.244	-0.024	-0.234	-0.172	0.062
	Sleep116	61.225	10	<0.001		-0.183	0.037		-0.209	0.025
	Sleep20	35.083	10	<0.001		-0.197	0.023		-0.261	-0.027
	Sleep44	39.024	10	<0.001		-0.246	-0.026		-0.284	-0.050
Ability to participate in social roles and activities	SRPPER11_CaPS	80.795	10	<0.001	0.127	0.151	0.024	-0.178	-0.140	0.038
	SRPPER18_CaPS	46.064	10	<0.001		0.111	-0.016		-0.180	-0.002
	SRPPER23_CaPS	43.280	10	<0.001		0.131	0.004		-0.199	-0.021
	SRPPER46_CaPS	20.431	10	0.025		0.123	-0.004		-0.186	-0.008
Pain Interference	PAININ9	19.672	10	0.032	0.058	0.064	0.006	0.190	0.212	0.022
	PAININ22	31.333	10	<0.001		0.063	0.005		0.198	0.008
	PAININ31	22.015	10	0.015		0.048	-0.010		0.188	-0.002
	PAININ34	40.246	10	<0.001		0.048	-0.010		0.177	-0.013

Table 3: Difference in latent variable means between the full invariant model and when the respective item thresholds and loadings are allowed to differ between groups. Differences are only displayed for the latent variable which the respective item loads on as other factor scores were not affected considerably (max Δ≤0.002)

			Percentile								
Domain	**Country**	**n**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
Physical Function	UK	1509	37.9 [36.6-39.1]	43.5 [42.5-44.5]	47.1 [46.2-48.0]	50.3 [49.4-51.1]	53.1 [52.2-54.0]	55.6 [54.8-56.5]	58.0 [57.3-58.8]	60.5 [59.8-61.3]	63.9 [63.1-64.7]
	France	1501	41.2 [40.2-42.2]	45.2 [44.4-46.0]	48.2 [47.4-49.0]	50.9 [50.1-51.7]	53.5 [52.7-54.4]	56.0 [55.2-56.8]	58.4 [57.7-59.1]	60.9 [60.2-61.7]	64.2 [63.3-65.2]
	Germany	1502	38.8 [37.8-39.8]	42.9 [42.1-43.6]	45.8 [45.1-46.6]	48.6 [47.7-49.4]	51.4 [50.6-52.3]	54.3 [53.4-55.1]	57.0 [56.3-57.8]	59.9 [59.1-60.7]	63.5 [62.4-64.5]
Anxiety	UK	1509	37.6 [36.6-38.6]	42.6 [41.4-43.8]	47.0 [45.9-48.0]	50.5 [49.6-51.4]	53.5 [52.7-54.3]	56.2 [55.5-56.9]	58.9 [58.1-59.7]	62.3 [61.5-63.2]	67.1 [66.0-68.3]
	France	1501	37.4 [36.3-38.5]	42.2 [41.2-43.3]	46.3 [45.3-47.3]	50.0 [49.1-50.9]	53.0 [52.3-53.8]	55.6 [55.0-56.3]	58.1 [57.5-58.8]	61.0 [60.3-61.8]	64.7 [63.8-65.6]
	Germany	1502	39.6 [38.3-41.0]	45.0 [43.9-46.2]	48.5 [47.8-49.3]	51.1 [50.4-51.7]	53.1 [52.5-53.7]	55.2 [54.6-55.8]	57.4 [56.8-57.9]	59.9 [59.2-60.7]	63.7 [62.9-64.6]
Depression	UK	1509	36.9 [35.9-38.0]	41.1 [40.2-42.0]	44.8 [43.8-45.9]	48.6 [47.6-49.7]	52.3 [51.3-53.2]	55.4 [54.6-56.2]	58.5 [57.6-59.4]	62.1 [61.3-63.0]	67.0 [66.0-68.1]
	France	1501	36.2 [35.2-37.2]	40.1 [39.2-41.0]	43.3 [42.3-44.3]	46.9 [45.9-47.8]	50.0 [49.1-50.9]	52.9 [52.1-53.6]	55.5 [54.8-56.2]	58.4 [57.6-59.1]	62.1 [61.2-63.0]
	Germany	1502	38.3 [37.0-39.5]	43.4 [42.4-44.5]	47.6 [46.7-48.5]	50.4 [49.7-51.1]	52.7 [52.0-53.3]	54.9 [54.2-55.5]	57.1 [56.5-57.8]	60.0 [59.3-60.6]	63.6 [62.7-64.4]
Fatigue	UK	1509	35.2 [33.9-36.5]	40.9 [39.8-41.9]	44.5 [43.7-45.3]	47.2 [46.5-47.8]	49.5 [48.8-50.2]	52.2 [51.4-53.0]	55.7 [54.7-56.6]	59.9 [58.8-61.1]	65.5 [64.5-66.5]
	France	1501	32.3 [31.4-33.2]	36.4 [35.5-37.4]	40.4 [39.4-41.4]	44.0 [43.1-44.9]	46.9 [46.1-47.6]	49.3 [48.6-50.0]	52.2 [51.4-53.1]	56.1 [55.1-57.1]	61.3 [60.2-62.5]
	Germany	1502	33.6 [32.5-34.8]	38.9 [37.9-39.9]	43.1 [42.2-44.0]	46.0 [45.3-46.8]	48.3 [47.7-49.0]	50.7 [50.0-51.4]	53.7 [52.8-54.5]	57.4 [56.5-58.3]	62.5 [61.5-63.6]
Sleep Disturbance	UK	1509	39.1 [38.0-40.1]	43.3 [42.5-44.1]	46.3 [45.5-47.1]	48.9 [48.2-49.6]	51.3 [50.7-52.0]	53.6 [52.9-54.3]	56.1 [55.3-56.8]	59.0 [58.2-59.7]	63.4 [62.3-64.4]
	France	1501	38.0 [37.0-38.9]	41.9 [41.1-42.6]	44.7 [44.0-45.5]	47.3 [46.6-48.0]	49.7 [49.1-50.4]	51.9 [51.3-52.6]	54.3 [53.5-55.0]	56.8 [56.1-57.5]	60.3 [59.5-61.2]
	Germany	1502	37.5 [36.6-38.5]	41.5 [40.8-42.3]	44.4 [43.6-45.1]	46.8 [46.1-47.6]	49.2 [48.5-50.0]	51.7 [51.0-52.4]	54.3 [53.6-55.0]	57.3 [56.4-58.1]	61.5 [60.6-62.5]
Ability to participate in social roles and activities	UK	1509	38.6 [37.6-39.6]	43.2 [42.4-44.1]	46.6 [45.7-47.5]	49.7 [49.0-50.5]	52.2 [51.5-52.9]	54.8 [54.0-55.6]	58.4 [57.4-59.5]	62.5 [61.5-63.5]	66.6 [65.6-67.6]
	France	1501	42.7 [41.9-43.6]	46.4 [45.6-47.1]	49.4 [48.7-50.0]	51.6 [51.0-52.2]	53.7 [53.1-54.3]	56.0 [55.3-56.8]	59.0 [58.2-59.9]	62.7 [61.7-63.6]	66.7 [65.8-67.5]
	Germany	1502	39.2 [38.4-40.0]	43.0 [42.2-43.7]	45.8 [45.1-46.6]	48.5 [47.9-49.2]	50.8 [50.2-51.4]	52.9 [52.3-53.5]	55.4 [54.6-56.1]	58.9 [57.9-59.9]	64.3 [63.1-65.4]
Pain Interference	UK	1509	36.4 [35.4-37.3]	40.0 [39.1-40.9]	43.0 [42.1-43.9]	46.3 [45.2-47.3]	49.8 [48.8-50.9]	53.1 [52.3-53.9]	55.7 [55.0-56.4]	59.1 [58.1-60.1]	64.9 [63.6-66.1]
	France	1501	36.4 [35.3-37.4]	40.1 [39.2-41.0]	43.2 [42.2-44.1]	46.4 [45.4-47.5]	50.0 [49.1-50.9]	52.8 [52.2-53.5]	55.1 [54.6-55.7]	57.7 [57.0-58.5]	61.7 [60.9-62.6]
	Germany	1502	37.2 [36.2-38.1]	41.1 [40.2-42.1]	44.9 [43.8-46.0]	49.0 [47.9-50.0]	52.4 [51.6-53.1]	54.6 [54.0-55.2]	56.7 [56.1-57.3]	59.9 [59.0-60.7]	64.4 [63.4-65.4]

Table 4: Country specific domain scores and the respective confidence interval for the 10th to 90th percentile of the general population

	0	1	2	3	4	5	6	7	8	9	10
UK	30.8 [28.5; 33.2]	48.0 [45.5; 50.6]	58.9 [56.4; 61.4]	68.3 [65.9; 70.5]	73.7 [71.5; 75.9]	79.9 [77.8; 81.8]	87.3 [85.5; 88.9]	93.3 [91.9; 94.4]	98.6 [97.9; 99.1]	99.7 [99.2; 99.9]	100.0 [99.8; 100.0]
France	24.3 [22.2; 26.6]	44.3 [41.8; 46.8]	59.6 [57.1; 62.1]	70.2 [67.8; 72.5]	76.7 [74.5; 78.8]	83.6 [81.6; 85.3]	90.4 [88.8; 91.8]	96.1 [95.0; 97.0]	98.8 [98.2; 99.3]	99.8 [99.4; 99.9]	100.0 [99.7; 100.0]
Germany	23.7 [21.6; 25.9]	45.5 [43.0; 48.0]	59.1 [56.6; 61.5]	69.6 [67.2; 71.8]	76.1 [73.9; 78.2]	83.1 [81.1; 84.9]	88.9 [87.2; 90.4]	94.7 [93.4; 95.7]	98.4 [97.7; 98.9]	99.6 [99.1; 99.8]	100.0 [99.7; 100.0]

Table 5: Cumulative percentages and the respective 95% confidence interval of pain intensity raw scores by country