Journal of Affective Disorders Reports 1 (2020) 100020

Contents lists available at ScienceDirect

Journal of Affective Disorders Reports

journal homepage: www.elsevier.com/locate/jadr

Research Paper

The WHO-5 well-being index – validation based on item response theory and the analysis of measurement invariance across 35 countries

Philipp E. Sischka[a,*], Andreia P. Costa[a], Georges Steffgen[a], Alexander F. Schmidt[b]

[a] Department of Behavioural and Cognitive Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[b] Institute of Psychology, Social & Legal Psychology, Johannes Gutenberg University Mainz

ARTICLE INFO

Keywords:
WHO-5
Item response theory
Measurement invariance
Differential item functioning
Cross-cultural research
Short scale
Well-being
Depression

ABSTRACT

Background: The five-item World Health Organization Well-Being Index (WHO-5) is a frequently used brief standard measure in large-scale cross-cultural clinical studies. Despite its frequent use, some psychometric questions remain that concern the choice of an adequate item response theory (IRT) model, the evaluation of reliability at important cutoff points, and most importantly the assessment of measurement invariance across countries.

Methods: Data from the 6th European Working Condition survey (2015) were used that collected nationally representative samples of employed and self-employed individuals (N = 43,469) via computer-aided personal interviews across 35 European countries. An in-depth IRT analysis was conducted for each country, testing different IRT assumptions (e.g., unidimensionality), comparing different IRT-models, and calculating reliabilities. Furthermore, measurement invariance analysis was conducted with the recently proposed alignment procedure.

Results: The graded response model fitted the data best for all countries. Furthermore, IRT assumptions were mostly fulfilled. The WHO-5 showed overall and at critical points high reliability. Measurement invariance analysis revealed metric invariance but discarded scalar invariance across countries. Analysis of the test characteristic curves of the aligned graded response model indicated low levels of differential test functioning at medium levels of the WHO-5, but differential test functioning increased at more extreme levels.

Limitations: The current study has no external criterion (e.g., structured clinical interviews) to assess sensitivity and specificity of the WHO-5 as a depression screening-tool.

Conclusions: The WHO-5 is a psychometrically sound measure. However, large-scale cross-cultural studies should employ a latent variable modeling approach that accounts for non-invariant parameters across countries (e.g., alignment).

1. Introduction

The World Health Organization has defined health since 1948 as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (World Health Organization, 2020; p. 1). Thus, subjective well-being is an important dimension of perceived quality of life. It encompasses negative aspects such as the presence of depression and anxiety but also positive aspects such as contentment, satisfaction, and happiness (McDowell, 2010). For these reasons, well-being is a common outcome measure across different populations (e.g., clinical vs. non-clinical) or as an indicator of treatment efficacy (e.g., pre-post treatment or comparisons of different treatment conditions).

1.1. The WHO-5 well-being index

Among numerous assessments of well-being in patient and non-patient populations the five-item World Health Organization Well-Being Index (WHO-5; World Health Organization, 1998; Topp et al., 2015) is one of the most widely used measures. The WHO-5 allows for a brief assessment (under 1 min) of well-being over a two-week period. Individuals are asked to indicate for each of the five statements how they felt over the past two weeks using a six-point Likert scale ranging from 0 = “at no time” to 5 = “all of the time”. The WHO-5 is derived from a 28-item version based on items from the Zung scales for depression, distress and anxiety as well as from the General Health Questionnaire and the Psychological General Well-Being Scale (Bech, 1993). Further analysis of these items resulted in an overall index of ten items (WHO-Ten) representing positive and negative well-being (Bech et al., 1996). A subsequent examination of the WHO-Ten revealed that five of the ten items focused on being interested in things. These items were there-

* Corresponding author.
E-mail address: philipp.sischka@uni.lu (P.E. Sischka).

https://doi.org/10.1016/j.jadr.2020.100020
Received 12 August 2020; Received in revised form 13 October 2020; Accepted 18 October 2020
Available online 24 October 2020

fore condensed into the item “My daily life has been filled with things that interest me”. A negative phrased item regarding depression was converted into a positively phrased item “I have felt cheerful and in good spirits”. The other three items remained the same (World Health Organization, 1998). The WHO-5 measures a global hedonic dimension of well-being (Bech, 2012).

The WHO-5 is most commonly used to assess clinical outcomes in controlled clinical trials and has shown to be a good measure of responsiveness/sensitivity to treatment (see Topp et al., 2015 for a review of the literature on the use of the WHO-5 in clinical trials). Although the WHO-5 was originally developed as a measure of well-being, it is also often applied as a screening tool for depression, showing high sensitivity for this condition (Topp et al., 2015). It represents aspects closely related to depression such as (lack of) positive mood, interests, and energy (Krieger et al., 2014). It has been adopted in various research fields such as suicidology (Sisask et al.,2008), geriatrics (Allgaier et al., 2013), youth problems (Rose et al., 2017), alcohol abuse (Elholm et al., 2011), diabetes (Halliday et al., 2017), stroke (Damsbo et al., 2020), cancer (Van Gestel et al., 2007), sleep (Hartwig et al., 2019), personality disorder (Fowler et al., 2018), grief (Killikelly et al., 2019), and occupational psychology (Sischka et al., 2020) among others. Moreover, the WHO-5 has been used in research projects all over the world (e.g., Topp et al., 2015) and also in large scale, multinational studies, involving different study fields such as diabetes (Nicolucci et al., 2013), quality of life (Eurofound (Ed.), 2017), quality of life in the actual coronavirus crisis (Eurofound (Ed.), 2020), and value research (Boye, 2009).

1.2. Psychometric properties of the WHO-5 – state of research

Most studies examining the psychometric properties focus on sensitivity and specificity of the WHO-5 to identify depression given an external criterion (e.g., structured clinical interviews; see Topp et al., 2015), reliability (usually investigated by calculating Cronbach’s α) and unidimensionality (e.g., through exploratory factor or principal component analyses). Despite the extensive use of the WHO-5, several important psychometric issues remain unclear (El-Den et al., 2018). This concerns the choice of an adequate item response theory (IRT) model, the reliability of the WHO-5, as well as the measurement invariance across countries.

1.3. Best-fitting item response theory model

The choice of an IRT model is an important aspect for item analysis as the different models have varying psychometric properties and rest on differential underlying assumptions (e.g., Ostini et al., 2014). Many studies on the psychometric properties of the WHO-5 apply the partial credit model (PCM; also called ordinal Rasch model; Masters, 1982) that assumes a constant discrimination parameter across items. The standard procedure to generate the WHO-5 – summing up the five items that yield theoretical raw scores from 0 (absence of well-being) to 25 (maximal well-being) – requires equal discrimination parameters.[1] However, IRT models that estimate separate discrimination parameters like the generalized partial credit model (GPCM; Muraki, 1992) or the graded response model (GRM; Samejima, 1969) might be more appropriate.

Unfortunately, to the best of the authors’ knowledge, so far, no thorough measurement model comparison has been conducted. Previous studies chose a measurement model based on (often not explicitly stated) assumptions of the measurement structure but neglected empirical data and model fit. The hitherto tested measurement models include Mokken analysis (e.g., Bech et al., 2003), PCM (e.g., Lucas-Carrasco et al., 2012), and confirmatory factor analysis (CFA) modeling the items as continuous (e.g., Saipanish et al., 2009) or categorical (e.g., Krieger et al., 2014). CFA with categorical data is equivalent to the (normal ogive) GRM.[2]

1.4. Reliability and standard error at specific cutoffs

Most of the studies, investigating the psychometric properties of the WHO-5 calculate Cronbach’s α as a measure of internal consistency (e.g., Krieger et al., 2014). Beside the fact that α has some problematic properties (e.g., Sijtsma, 2009), it is based on the assumption that the standard error of measurement is uniform across the latent variable continuum, i.e., it is an empirical estimate of a measure’s marginal reliability. Within IRT models this assumption can be tested with test information functions. Reliability at a given point of the test information function can be calculated as: reliability = 1 – 1/test information, whereas the standard error can be calculated as: SE = 1 / $\sqrt{\text{test information}}$ (Brown, 2018). The issue of non-uniform measurement standard error is especially important for the WHO-5, as it is often used as a screening tool for depression with cutoffs of ≤ 28 and ≤ 50 on the 0-100 scale (e.g., Topp et al., 2015; Löwe et al., 2004; Nicolucci et al., 2013). The open empirical question, however, is whether the WHO-5 has acceptable reliability at these cutoff points.

1.5. Measurement invariance

A further issue concerns the comparability of the WHO-5 across countries as many multinational studies use it to compare the well-being across countries. The importance to guarantee the comparability of theoretical constructs over the compared units (e.g., countries) has been emphasized (e.g., Harkness et al., 2003; Millsap, 2011). The measurement structures of the latent construct and their corresponding manifest items need to be (at least partially) stable across the compared research units in order to be comparable. Therefore, measurement invariance (MI) testing is a necessary precondition for comparative analyses (Millsap, 2011). If MI is not tested, (the lack of) differences between groups on the latent or manifest constructs cannot be unambiguously attributed to ‘real’ (non-)differences as these could also be caused by differences in the measurement attributes (e.g., Whittaker, 2013). Hence, non-invariance (across countries) could emerge if (a) the conceptual meaning or understanding of the construct differs across groups, if (b) groups differ regarding the extent of social desirability or social norms, if (c) groups have different reference points, when making statements about themselves, if (d) groups respond to extreme items differently, if (e) particular items are more applicable for one group than another, or if (f) the translation of one or more items is improper (Chen, 2008). It is worth noting that full MI (i.e., items do not show differential item functioning across groups) is not strictly necessary to compare latent means across groups. Instead, partial invariance (i.e., at least some items show no differential item functioning across groups) is sufficient for such comparisons (Reise et al., 1993). However, the more items do not show differential item functioning across groups, the more reliably the means are estimated (Steenkamp & Baumgartner, 1998).

As certain aspects of the WHO-5 psychometric properties hitherto remain untested (i.e., lack of model comparison, reliability testing, MI across countries) we conducted an in-depth psychometric analysis of the WHO-5 in a large sample (43,469 respondents) of 35 European countries within an IRT framework. We tested crucial assumptions of IRT modeling (i.e., unidimensionality, local independence, monotonicity) and compared different IRT models (i.e., PCM, GPCM, GRM) against

[1] Many studies, for whatever reason, multiply the raw score by four to obtain a score ranging from 0 to 100 in which 0 represents the worst and 100 the best imaginable well-being (Topp et al., 2015).

[2] However, in practice CFA with categorical data is most often estimated with limited information estimators (e.g., WLSMV) that only use a summary of the available data (i.e., variances and covariances), whereas GRM is most often estimated with full information estimators (e.g., full-information maximum likelihood) that use the raw data (Edwards et al., 2012; Wirth & Edwards, 2007).

each other. Furthermore, we investigated psychometric properties of the WHO-5 and analyzed whether it has sufficient reliability at frequently-used cutoff values. Moreover, we tested the WHO-5 for MI (or differential test functioning; DTF) across countries. Moreover, we sought to identify non-invariant parameters across countries using the recently proposed alignment method (Asparouhov & Muthén, 2014; Marsh et al., 2018). Thus, we aimed to answer recent calls to establish cross-cultural MI for established measures (Boer et al., 2018).

2. Methods

2.1. Survey design and participants

We used publicly accessible data from the European Working Condition Survey 2015 (European Foundation for the Improvement of Living and Working Conditions, 2017; for survey details see Eurofound, 2015a). This survey assessed and quantified working conditions of employed and self-employed individuals across Europe within nationally representative samples. The target population was comprised of residents who had worked for pay or profit for at least one hour in the week preceding the interview. The survey covered the EU28 countries as well as Norway, Switzerland, Albania, the former Yugoslav Republic of Macedonia, Montenegro, Serbia, and Turkey between February and September 2015 (December 2015 in Albania, the former Yugoslav Republic of Macedonia, Montenegro, Serbia, and Turkey). The survey was conducted in all countries via computer-aided personal interviewing at respondents' homes. The sample selection was based on a multi-stage process resulting in a complex survey sampling. Therefore, a weighting variable was used that accounts for unequal sample selection probability and adjusts the sample so that it reflects the socio-demographic structure of the target population (post-stratification, Eurofound, 2015b).

The initial sample consisted of N = 43,850 respondents. Due to incomplete data (i.e., one or more missing values on the WHO-5 items), 0.9% (n = 381) of participants had to be excluded from the analyses. Therefore, the effective sample consisted of 43,469 respondents (49.6% females, n = 21,553, number of respondents ranged between 946 and 3346 per country). The interviewees' age ranged from 15 to 89 years (M = 43.3, SD = 12.7; see Table A1 in the Electronical supplement for further sample details).

2.2. Measures

All five items of the WHO-5 were included. English items are (1) "I have felt cheerful and in good spirits.", (2) "I have felt calm and relaxed.", (3) "I have felt active and vigorous.", (4) "I woke up feeling fresh and rested.", (5) "My daily life has been filled with things that interest me." The instruction read "Please indicate for each of the five statements which is the closest to how you have been feeling over the last two weeks". Response options ranged from 0 (*At no time*) to 5 (*All of the time*).

2.3. Statistical analysis

To test the unidimensionality assumption required for the application of IRT models, parallel analysis (Horn, 1965) and minimum average partial method (Velicer, 1976) with principal component analysis and polychoric correlations were used that have been found to work well with ordinal variables (Garrido et al., 2011, 2013). In a next step, the different IRT models (i.e., PCM, GPCM, GRM) were compared using the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) as well as their sample size adjusted variants to determine which model fitted the data best. However, if these information criteria disagreed, the BIC was used as the final criterion as it has been found to be most accurate in selecting the correct polytomous IRT model (Kang et al., 2009). Moreover, the Vuong test (Vuong, 1989) was used that has been found to reliably distinguish between the GRM and GPCM (Schneider et al., 2019). It first tests, whether two models are distinguishable. In a second step it ascertains whether the second model shows better fit than the first model. Additionally, and in order to assess the relative improvement in the proportion of variability accounted for by one model over the other, we calculated the R^2 based on the likelihood ratio G^2 test statistic that can be regarded as analogous to comparing various regression models' R^2s (De Ayala, 2009).

The accuracy of the selected IRT model was further analyzed with goodness of fit statistics relying on the limited-information test statistic C_2 that was specifically developed for IRT models with only few items (Cai & Monroe, 2014; Monroe & Cai, 2015). We calculated the *RMSEA* (acceptable values ≤.08, good ≤.05), *SRMSR* (acceptable ≤ .1, good ≤ .08), *TLI* and *CFI* (both acceptable ≥ .90, good ≥ .95) according to commonly used thresholds (Kline, 2015). However, as goodness of fit statistics based on the limited-information test statistic are relatively new, caution should be exercised in interpreting these statistics on simple rules of thumb (Cai & Hansen, 2013; Cai & Monroe, 2013; Monroe & Cai, 2015; Maydeu-Olivares & Joe, 2014). It should also be noted that the *RMSEA* based on C_2 is influenced by the number of answer categories of the items as more answer categories tend to increase *RMSEA* values (Monroe & Cai, 2015). Furthermore, item and test characteristic curves (ICC, TCC) were generated. Additionally, item and test information functions (IIF, TIF) were derived and empirical marginal reliability as well as root mean square standard error (*RMSE*)[3] were calculated as summary measures of score precision (Brown, 2018; DeMars, 2018). Moreover, the Jackknife Slope Index (JSI; Edwards et al., 2018) was used to evaluate the assumption of local independency. The JSI is based on the phenomenon that locally dependent items will lead to artificially inflated slope parameters. Positive JSI values indicate that the removal of a specific item caused the slope of another item to decrease, while negative JSI values indicate that the removal of a specific item lead to an increased slope of another item. The proposed cutoff values by Edwards et al. (2018; mean of the JSI values plus twice the standard deviation) were used as indication of local dependency. Then, the generalized S-X^2 item fit index (Kang and Chen, 2011) and corresponding *RMSEA* as measure of effect size were calculated to assess item fit. Finally, possible violations from the assumption of monotonicity with raw residual plots were inspected (Wells and Hambleton, 2016).

In the next step we tested the WHO-5 for MI across countries. To this end, we conducted a multigroup analysis with increasingly restrictive nested models. In the *configural model* (the baseline model) in all countries the factor mean was fixed to zero, the factor variance was fixed to one and the discrimination and difficulty parameters were freely estimated. The next more restrictive model was the *metric invariance* model that constrained the discrimination parameter to be equal across all countries, while the factor variance was fixed to one only in the first group. In this model the factor mean was still fixed to zero and the difficulty parameter was freely estimated in all countries. The most restrictive model was the *scalar invariance model* that constrained the discrimination parameter as well as the difficulty parameter to be equal, while the factor mean was fixed to zero and the factor variance was fixed to one only in the reference group. Factor mean and factor variance were freely estimated in all other groups.

Finally, we used the alignment optimization method to find non-invariant parameters (Asparouhov and Muthén, 2014; Marsh et al., 2018). The alignment method was originally developed within a confirmatory factor analysis framework (Asparouhov and Muthén, 2014), but was quickly adopted in an IRT framework (Muthén and Asparouhov, 2014) with promising results from further simulation studies

[3] The formula 20.22 in Brown (2018) was used and the root of the mean squared standard error was calculated.

(DeMars, 2020; Finch, 2016). Based on a simplicity function (similar to the rotation criteria used with exploratory factor analysis), this method searches for the most optimal MI model with the configural invariance model as baseline model. It finds the strongest source of non-invariance in a minimal number of items, while allowing the other majority of items to have a trivial amount of differences (Kim et al., 2017). Therefore, the aligned model has the same model fit as the configural invariance model (Asparouhov and Muthén, 2014). As the algorithm is based on multiple testing, Asparouhov and Muthén (2014) used significance level of .01 to determine a starting set of invariant groups. After the determination of the invariance set, a significance test (significance level of .001) is conducted to compare the parameter value for each group with the parameter average of the invariant groups. The alignment method provides an R^2 value that represents the parameter variation across groups in the configural model that is explained by variation in the factor mean and factor variance across groups. A value close to one indicates a high degree of invariance, while a value close to zero indicates a low degree of invariance (Asparouhov and Muthén, 2014).

Two alignment optimization methods can be used: In the FIXED approach, the factor mean and factor variance of a reference group is set to 0 and 1, respectively. Typically, the group with factor mean closest to 0 is used as reference group, to avoid misspecification and estimation biases. In the FREE approach there is no constraint on the reference group's factor mean, and it is freely estimated (Asparouhov and Muthén, 2014). Asparouhov and Muthén (2014) recommend using the FREE approach when more than two groups are being compared and when measurement noninvariance exists. Guidance on the implementation of the alignment method and technical details such as the computation of the loss function are described in Asparouhov and Muthén (2014, see also Kim et al., 2017 and Marsh et al., 2018).

The reliability of the alignment method depends on the precision of the factor mean and variance estimation. Possible problems with factor mean and variance estimations, were checked with a simulation study with 500 simulation runs (Muthén and Asparouhov, 2014). A near-perfect correlation for the ordering of countries with respect to factors to be trustworthy is required. Muthén and Asparouhov (2014) recommend a correlation of at least .98. Moreover, relative parameter bias (defined as = [alignment parameter – average of the parameter estimates across replications] / alignment parameter * 100) for the means and the proportion of replications for which the 95% confidence interval contains the mean were calculated.

The alignment method "serves the joint purposes of scale linking and purification, without literally deleting items from the linking" (DeMars, 2020, p. 56). Thus, the alignment method does not need anchor items for the scale linking between the compared groups that can be hard to identify if there is no prior knowledge about DIF-free items (Huelmann et al., 2020). Based on the IRT parameters from the alignment method, we investigated country pairwise DTF. The compensatory and non-compensatory differential response functioning (DRF) statistics were calculated as they represent suitable and interpretable effect sizes for response bias across groups (Chalmers, 2018). A compensatory DRF (sDRF) that is significantly different from zero indicates that one group, on average, receives higher scores on the test than the other group on the same level of the latent variable. However, if the response functions cross at one or more locations (i.e., one group gets higher values at certain levels of the latent variable and the other group at other levels), the compensatory DRF may approach the value of zero (Chalmers, 2018). This is often viewed as negative by psychometricians, because it is often more important to quantify the overall response bias rather than the degree of response bias after allowing for cancelation across the range of the latent variable (Chalmers, 2018). The non-compensatory DRF (uDRF), on the other hand, quantifies the overall response bias. The sDRF and uDRF are equal if the response functions of the two compared groups never cross.

3. Results

3.1. Descriptives

The items showed no extreme skewness and kurtosis (M_{skewness} = -0.84, SD_{skewness} = 0.25; M_{kurtosis} = 0.38, SD_{kurtosis} = 0.72; see Table A2 and Figure A1 in the Electronical supplement). The polychoric correlations between the items ranged between .38 and .87. Denmark showed the lowest item intercorrelations (M_{polycor} = .54, SD_{polycor} = .10) and Slovakia the highest (M_{polycor} = .82, SD_{polycor} = .03, see Figure A2 in the Electronical supplement).

3.2. Assessing unidimensionality, model comparison, and model-data fit

Results of parallel analysis and minimum average partial method revealed one dominant factor for all countries. Parallel analysis showed for each country that only for the first factor, the eigenvalue of the real data is greater than the eigenvalue from the random data. Furthermore, the minimum average partial method indicated that for each country only one factor should be extracted. Moreover, for every country, only one factor showed an eigenvalue greater than 1 that explained between 54% and 82% (see Figure A3 and A4 in the Electronical supplement for a detailed analysis for each country), corroborating the unidimensionality of the WHO-5.

Fig. 1 shows the different information criteria for the three models. Every criterion favored the GRM over the PCM and the GPCM for each country. Furthermore, the Vuong test revealed that the GRM and the GPCM were distinguishable and that the test favored the GRM over GPCM for each country (see Table A3 in the Electronical supplement). The ΔR^2 between the PCM and the GPCM for each country ranged between .112 and .288 and should be considered as a non-trivial improvement in every case. Comparing the GPCM and the GRM ΔR^2 ranged between .005 and .027, thus, the improvement was only small. Nevertheless, as the GPCM and GRM did not differ in terms of complexity, and because the Vuong test favored the GRM, the WHO-5 items were further analyzed based on the GRM.

Table 1 shows the goodness of fit statistics for the GRM model for each country. The C_2 statistic indicates close fit for Latvia, North Macedonia, Romania, and Slovenia. However, given the large sample sizes, even trivial model specification can lead to a significant C_2 statistic. The approximate fit indices *SRMSR* and *CFI* indicated good model fit for all countries. The *TLI* yielded at least acceptable model fit for all countries and good model fit for most countries (except for Denmark and France). However, the *RMSEA* revealed that the unidimensional GRM model does not fit well for some countries. Especially Belgium, Denmark, Spain, Finland, France, Lithuania, Luxembourg, and Montenegro had *RMSEA* values above .1.

The JSI flagged local dependence between item 3 and 4 for Switzerland, Luxembourg, Malta and Romania (see Fig. A5 in the Electronical supplement). For Montenegro, items 1 and 2 were flagged as being locally dependent. However, the values were only slightly above the threshold, thus, local dependency might be considered still in an acceptable range. The generalized S-X^2 item fit index flagged a lot of items to deviate from the GRM curves. However, the item-level *RMSEA* ranged between .000 and .050 for item 1, between .008 and .052 for item 2, between .000 and .066 for item 3, between .000 and .053 for item 4, and between .007 and .073 for item 5, indicating low to medium deviation of the items from the GRM (see Fig. A6 in the Electronical supplement). Finally, the raw residual plots indicated no strong deviation from monotonicity (see Fig. A7.1-A7.35 in the Electronical supplement).

3.3. Psychometric properties

Fig. 2 shows the item parameter for the GRM for each country. Items 1, 2, and 3 yielded on average higher discrimination parameters ($M_{\text{Item 1}}$ = 3.25, $SD_{\text{Item 1}}$ = 0.51; $M_{\text{Item 2}}$ = 2.97, $SD_{\text{Item 2}}$ = 0.64;

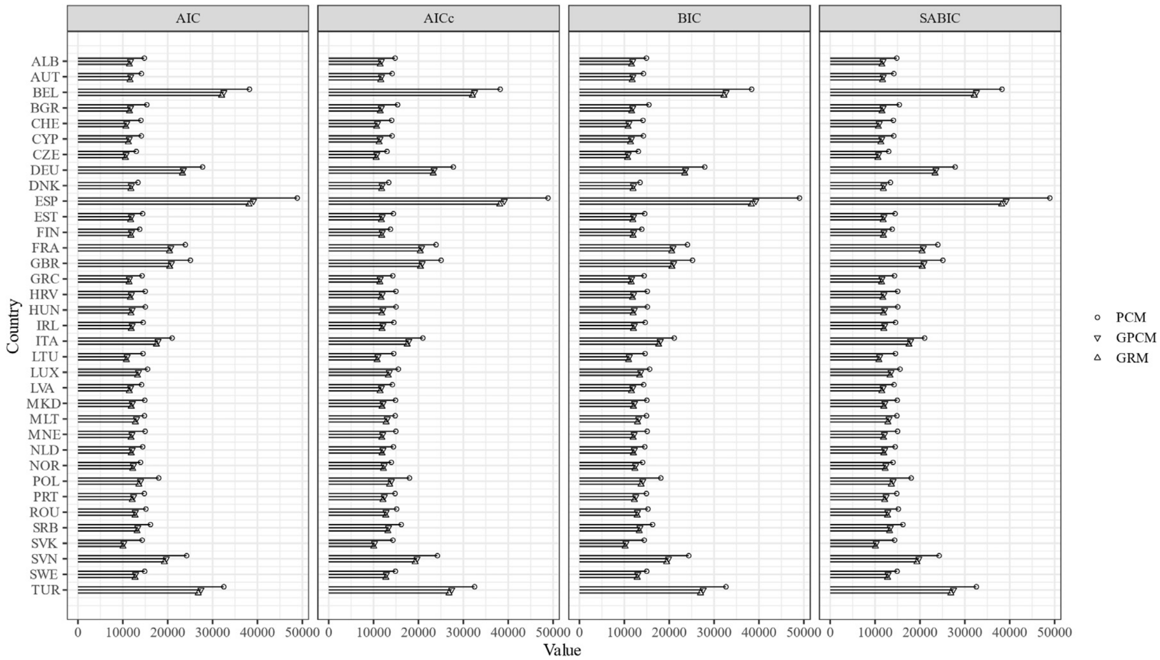

Fig. 1. Model comparison for the PCM, GPCM, and GRM.

Table 1
Goodness of fit statistics for the graded response model.

Country	C_2	*p*	*RMSEA* [90% CI]	*SRMSR*	*TLI*	*CFI*
ALB	50.505	0.000	.096 [.073; .120]	.028	.980	.990
AUT	29.836	0.000	.070 [.047; .095]	.030	.986	.993
BEL	151.001	0.000	.106 [.092; .121]	.053	.964	.982
BGR	40.224	0.000	.082 [.059; .106]	.026	.987	.994
CHE	35.047	0.000	.077 [.054; .103]	.028	.987	.993
CYP	30.648	0.000	.072 [.049; .097]	.032	.987	.993
CZE	52.756	0.000	.098 [.075; .123]	.034	.972	.986
DEU	59.584	0.000	.072 [.057; .089]	.030	.985	.992
DNK	79.361	0.000	.122 [.099; .146]	.055	.938	.969
ESP	315.005	0.000	.136 [.124; .149]	.063	.956	.978
EST	48.823	0.000	.094 [.071; .118]	.058	.973	.987
FIN	57.488	0.000	.103 [.080; .127]	.048	.959	.980
FRA	209.224	0.000	.164 [.145; .183]	.056	.915	.958
GBR	57.136	0.000	.080 [.062; .100]	.037	.983	.992
GRC	41.153	0.000	.085 [.062; .110]	.023	.982	.991
HRV	45.724	0.000	.090 [.067; .115]	.031	.982	.991
HUN	51.894	0.000	.096 [.073; .121]	.038	.978	.989
IRL	27.075	0.000	.065 [.042; .090]	.036	.988	.994
ITA	54.531	0.000	.084 [.065; .105]	.038	.980	.990
LTU	64.476	0.000	.110 [.087; .134]	.026	.977	.988
LUX	60.938	0.000	.106 [.083; .131]	.041	.962	.981
LVA	16.822	0.005	.050 [.025; .077]	.024	.994	.997
MKD	6.948	0.225	.020 [.000; .051]	.028	.999	1.000
MLT	49.474	0.000	.094 [.071; .119]	.051	.968	.984
MNE	82.955	0.000	.125 [.102; .149]	.030	.965	.982
NLD	42.307	0.000	.085 [.063; .110]	.031	.980	.990
NOR	43.944	0.000	.087 [.065; .112]	.051	.966	.983
POL	49.694	0.000	.087 [.066; .110]	.047	.984	.992
PRT	27.880	0.000	.067 [.044; .093]	.041	.987	.994
ROU	4.304	0.507	.000 [.000; .040]	.019	1.000	1.000
SRB	31.180	0.000	.071 [.049; .096]	.024	.988	.994
SVK	39.423	0.000	.085 [.061; .110]	.036	.987	.993
SVN	7.294	0.200	.017 [.000; .042]	.025	.999	1.000
SWE	50.692	0.000	.096 [.073; .120]	.036	.970	.985
TUR	59.262	0.000	.074 [.058; .091]	.074	.984	.992

Notes. df = 5; *RMSEA* = root mean squared error of approximation; *SRMR* = standardized root mean square residual; *TLI* = Tucker-Lewis index; *CFI* = comparative fit index.

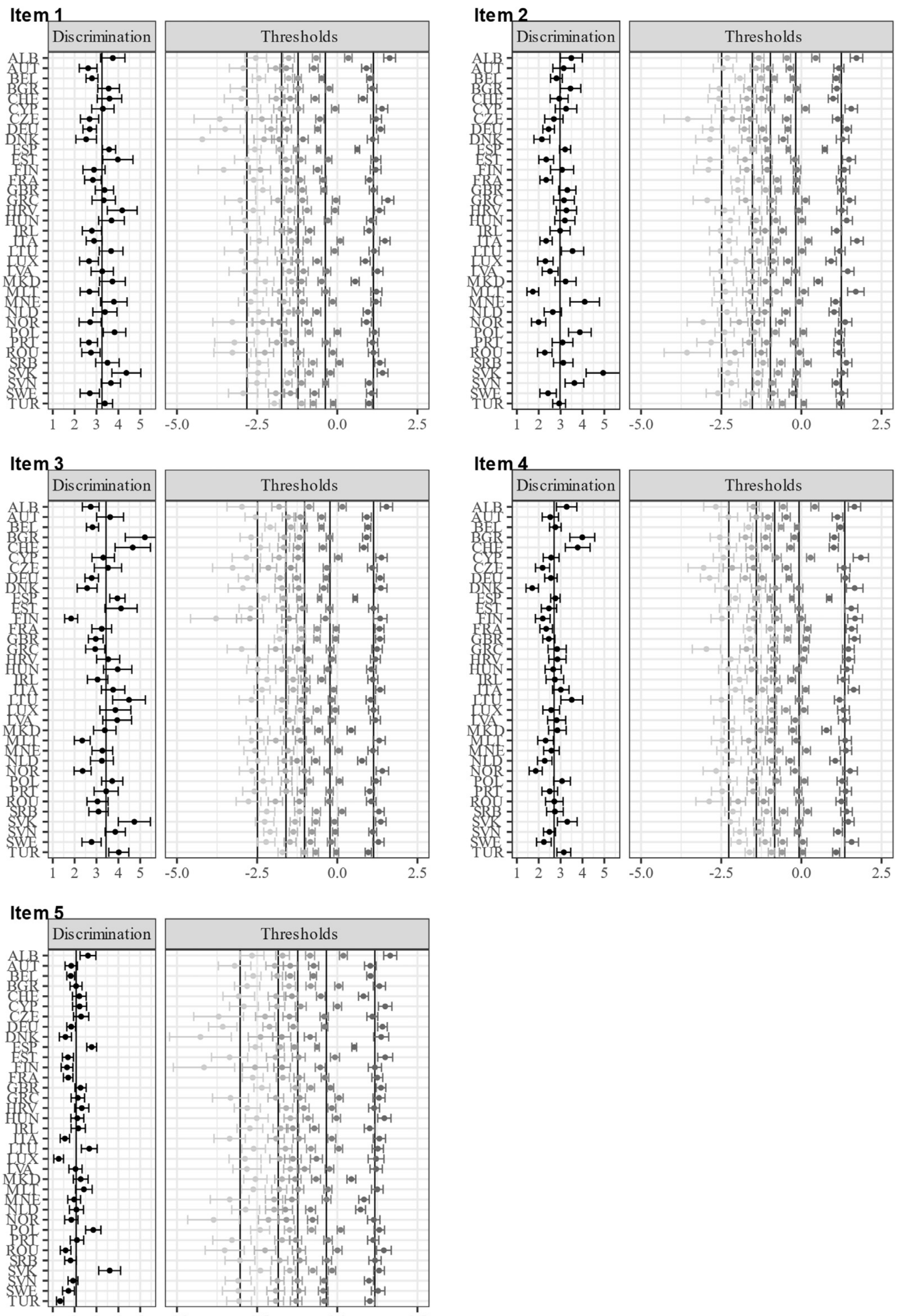

Fig. 2. Item parameter for the GRM.

$M_{\text{Item 3}} = 3.44$; $SD_{\text{Item 1}} = 0.73$), while items 4 and 5 revealed lower discrimination parameters ($M_{\text{Item 4}} = 2.70$, $SD_{\text{Item 1}} = 0.48$; $M_{\text{Item 5}} = 2.07$, $SD_{\text{Item 2}} = 0.46$). The items differed only slightly on item difficulty. Item characteristic curves and item information functions can be seen in Fig. A8 and A9 in the Electronical supplement.

Fig. 3 shows the item and test information functions. The empirical marginal reliability ranged between .83 and .93 (*RMSE* ranged between 0.257 and 0.412). The reliability was also satisfying at both frequently used cutoff values for depression, i.e., $\rho_{12.5}$ (WHO-5 score of 12.5 equals 50 on the 0-100 scale) ranged between .86 and .96 (standard errors ranged between 0.202 and 0.379) and ρ_7 (WHO-5 score of 7 equals 28 on the 0-100 scale) ranged between .84 and .95 (standard errors ranged between 0.212 and 0.398; see Table A6 in the Electronical supplement for each country).

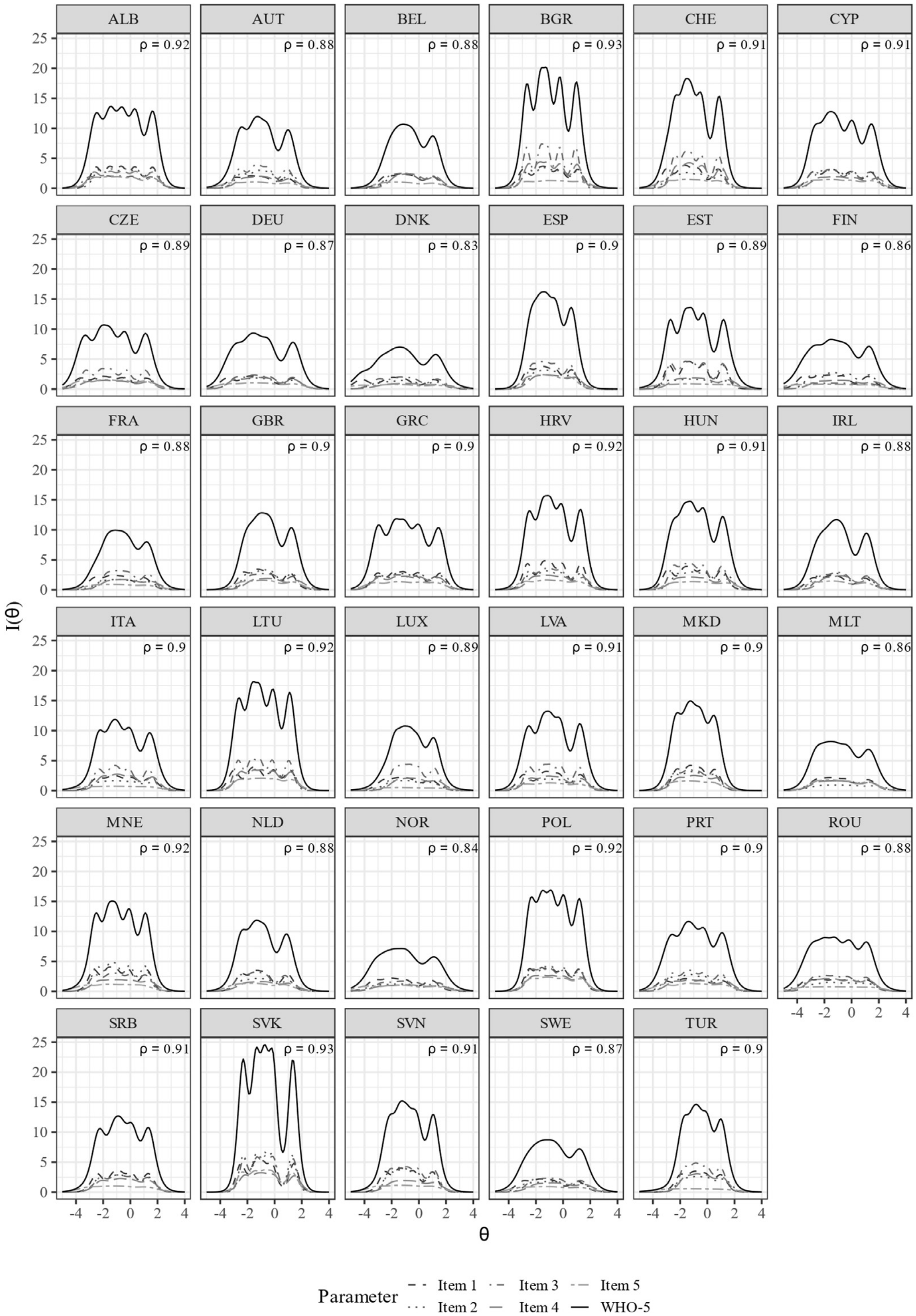

Fig. 3. Item and test information functions for the GRM.

3.4. Measurement invariance and differential test functioning analysis

The multigroup analysis (with the GRM as base model) revealed a very good fit for the configural model (C_2 = 2,073.244; *df* = 175, *p* = .000, *RMSEA* = .016, *SRMSR* for each country ranged between .019 and .074, *TLI* = .977, *CFI* = .988). The metric invariance model showed also a very good fit (C_2 = 3,248.038; *df* = 311, *p* = .000, *RMSEA* = .015, *SRMSR* ranged between .026 and .131, *TLI* = .980, *CFI* = .982), indicating that the item discrimination parameters between the countries were approximately invariant. However, fixing the item thresholds to be equal across countries led to strong deterioration of some model fit indicators (C_2 = 17,738.548; *df* = 1127, *p* = .000, *RMSEA* = .018, *SRMSR* ranged between .026 and .188, *TLI* = .968, *CFI* = .898), indicating non-invariance for at least some threshold parameters (see Table A7 in the Electronical Supplement for all *SRMSR* values).

Table 2 shows the fit statistics of the alignment analysis with the FREE approach and with Great Britain as reference group (with variance fixed to one). The average invariance index equaled .362 and 28.9% of

Table 2
Alignment fit statistics.

Item	Parameter	R^2	Weighted Average across invariant groups	Weighted Variance across invariant groups	Weighted Average across all groups	Weighted Variance across all groups	Number (percentage) of approx. invariant groups
Item 1	Discrimination	.728	3.28	0.24	3.28	0.24	35 (100%)
	Threshold 1	.340	−2.14	0.26	−1.99	0.32	26 (74.3%)
	Threshold 2	.539	−1.01	0.12	−0.95	0.16	28 (80%)
	Threshold 3	.626	−0.46	0.10	−0.46	0.12	26 (74.3%)
	Threshold 4	.071	0.61	0.10	0.36	0.25	14 (40%)
	Threshold 5	.000	1.86	0.24	1.80	0.33	27 (77.1%)
Item 2	Discrimination	.706	3.01	0.30	2.99	0.33	34 (97.1%)
	Threshold 1	.274	−1.84	0.25	−1.65	0.34	23 (65.7%)
	Threshold 2	.444	−0.89	0.13	−0.75	0.16	18 (51.4%)
	Threshold 3	.693	−0.22	0.07	−0.22	0.08	33 (94.3%)
	Threshold 4	.165	0.68	0.11	0.55	0.20	17 (48.6%)
	Threshold 5	.000	2.05	0.21	1.91	0.30	21 (60%)
Item 3	Discrimination	.532	3.53	0.39	3.48	0.46	33 (94.3%)
	Threshold 1	.331	−1.99	0.34	−1.67	0.40	18 (51.4%)
	Threshold 2	.500	−0.99	0.23	−0.83	0.24	18 (51.4%)
	Threshold 3	.603	−0.33	0.08	−0.28	0.12	22 (62.9%)
	Threshold 4	.400	0.59	0.13	0.49	0.17	22 (62.9%)
	Threshold 5	.000	1.81	0.22	1.80	0.28	29 (82.9%)
Item 4	Discrimination	.623	2.75	0.29	2.75	0.29	35 (100%)
	Threshold 1	.210	−1.77	0.25	−1.46	0.40	17 (48.6%)
	Threshold 2	.337	−0.78	0.13	−0.64	0.25	17 (48.6%)
	Threshold 3	.552	−0.17	0.06	−0.10	0.13	21 (60%)
	Threshold 4	.489	0.59	0.09	0.64	0.15	25 (71.4%)
	Threshold 5	.000	2.20	0.24	2.01	0.28	21 (60%)
Item 5	Discrimination	.517	2.18	0.26	2.10	0.34	31 (88.6%)
	Threshold 1	.354	−2.22	0.32	−2.20	0.34	31 (88.6%)
	Threshold 2	.415	−1.08	0.18	−1.07	0.19	29 (82.9%)
	Threshold 3	.410	−0.52	0.15	−0.48	0.19	28 (80%)
	Threshold 4	.000	0.26	0.12	0.37	0.30	16 (45.7%)
	Threshold 5	.000	1.85	0.26	1.84	0.33	21 (60%)

Notes. MLR estimator; FREE approach.

the parameters were non-invariant. The R^2 values for the item discrimination parameter ranged between .517 and .728 corroborating the finding of the multigroup analysis that the item discrimination parameters, with few exceptions were invariant across countries (the number of invariant discrimination parameters ranged between 31 and 35). Regarding the thresholds, the R^2 values were highest for the thresholds 2 and 3 for nearly all items, while items 1 and 4 showed lower R^2 values. The threshold 5 showed an R^2 value of .000 for all items. The standard deviation of the parameters showed the same pattern, i.e., it is often lower for the thresholds 2 and 3 and higher for the thresholds 1, 4, and 5. Strikingly, the number of approximate invariant groups showed a less clear picture (for a quick overview of invariant and non-invariant parameters see Fig. A11 in the Electronical Supplement). The simulation study revealed that the correlation between the population factor means and the estimated alignment factor means computed over groups and averaged over replications equals .98. Relative parameter bias for the factor means ranged between 2.4% and 9.6% (M = 4.9; SD = 1.5); the proportion of replications for which the 95% confidence interval contains the mean ranged between 89.0% and 96.8% (M = 93.2; SD = 1.5; see Table A8 in the Electronical Supplement for all information). Thus, factor mean estimates and factor mean rankings of the alignment method emerged as pretty reliable (Muthén and Asparouhov, 2014).

Fig. 4 shows the item parameter of the alignment procedure. Because of the scale linking of the alignment method, these parameters can be directly compared. Fig. 4 corroborates the finding that the discrimination parameters did not differ significantly and that the thresholds 2 and 3 showed less variation between countries. Fig. 4 also gives an indication of the ambiguous results of R^2 and variation of the parameters on the one hand, and the number of approximate invariant groups on the other hand. It shows that the standard error was much larger for the thresholds 1 and 5, thus, the power to establish significant non-invariance was lower for these parameters. These results indicate that there was less DIF and DTF in the middle range of the latent variable.

Fig. 5 shows the test characteristic curves of the WHO-5 for the GRM after alignment. As an illustration, it can be seen that the expected test score for negative values of the latent variable was especially high for Albania, Bulgaria, and Slovakia. That means that these countries had higher values as one would expect, given their latent variable level. On the other hand, these countries had lower expected test scores for positive values of the latent variable. Fig. 6 shows the sDRF and uDRF statistics for Great Britain as reference group (for the other country comparisons see Fig. A12.1-A12.35 in the Electronical supplement). The curves show DTF at different levels of the latent variable. Negative values indicated that the other group had higher expected test scores, whereas positive values indicate that the reference group (Great Britain) had higher expected test scores. The sDRF and uDRF statistics summarize the DTF across the full range of the latent variable. For instance, only minor DTF effects appeared between Great Britain and Austria, whereas larger DTF effects showed up between Great Britain and Albania. Interestingly, most of the DTF effects indicate unidirectional response bias, i.e., that the response functions of two countries cross at one or more locations (Chalmers, 2018). Considering all country comparisons, the sDRF statistics ranged between -1.51 and 1.36 (M = -0.07, SD = 0.57) and the uDRF statistics between 0.05 and 2.12 (M = 0.73, SD = 0.45; for a quick overview of the sDRF and uDRF statistics see Fig. A13 in the Electronical supplement).

3.5. Comparing alignment factor scores and manifest sum scores

The correlations between factor scores and manifest sum scores were quite high and ranged between .95 and .99 within each country (see Fig. A14 in the Electronical supplement). However, it could be observed that the regression slopes differed between countries and ranged between

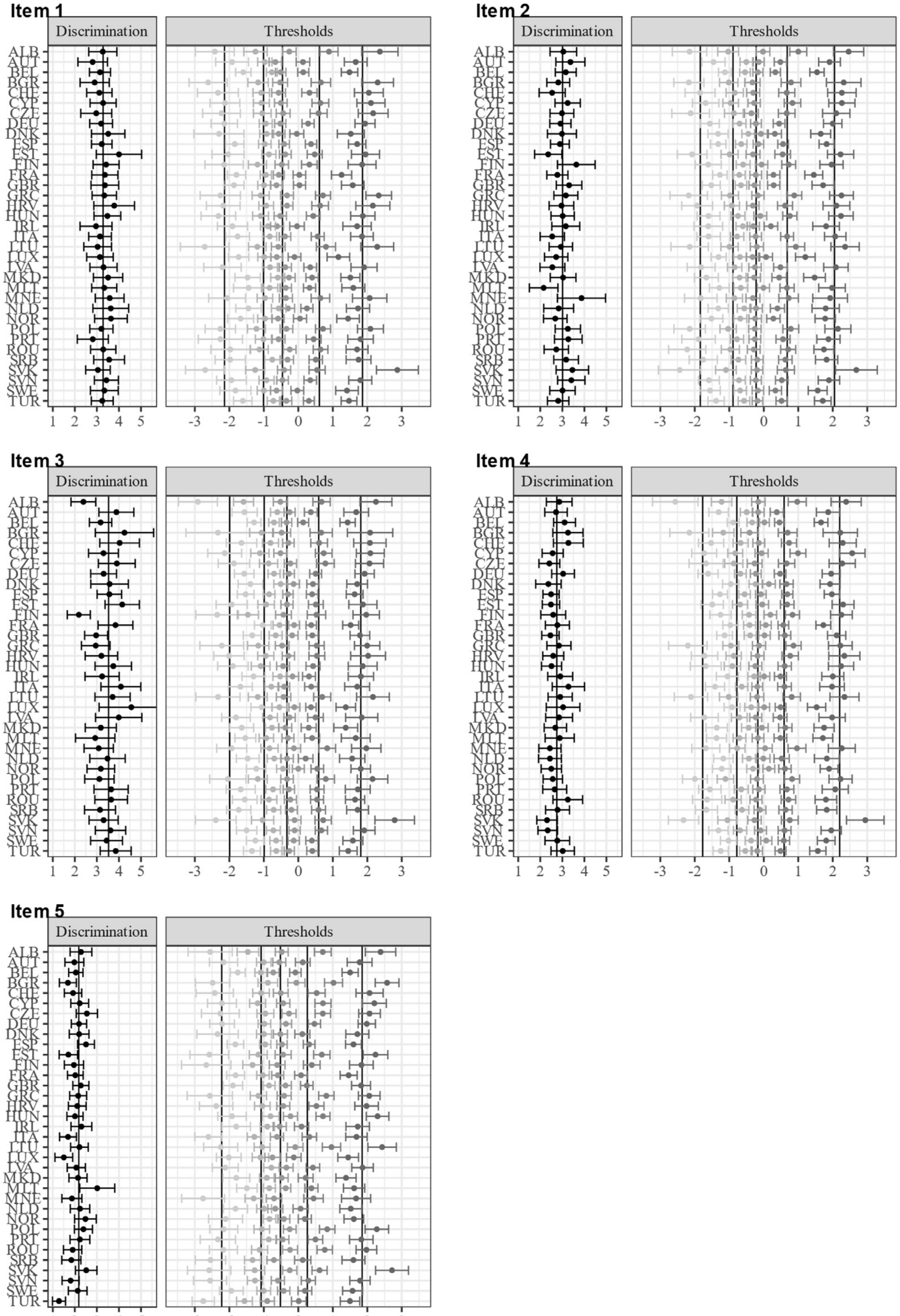

Fig. 4. Item parameter for the GRM after alignment.

3.8 and 6.6 reflecting the different test characteristic curves. Thus, the manifest scores and the factor scores differed also in their distributions. The correlation between the means of the manifest sum score and the means of the factor scores was only .77 (see Fig. 7). Thus, there was a non-negligible change in the country rank order regarding their mean well-being levels, when switching from manifest to latent models (see also Fig. A15 in the Electronical supplement). For instance, comparing manifest means, Luxembourg had nearly the same mean level of well-being as Lithuania. However, comparing the latent means, Lithuania had a significantly higher mean.

4. Discussion

The IRT analyses clearly revealed that the GRM provided better fit to the data than the PCM and the GPCM. This means that the items do not only differ in terms of difficulty but also in terms of discrimination.

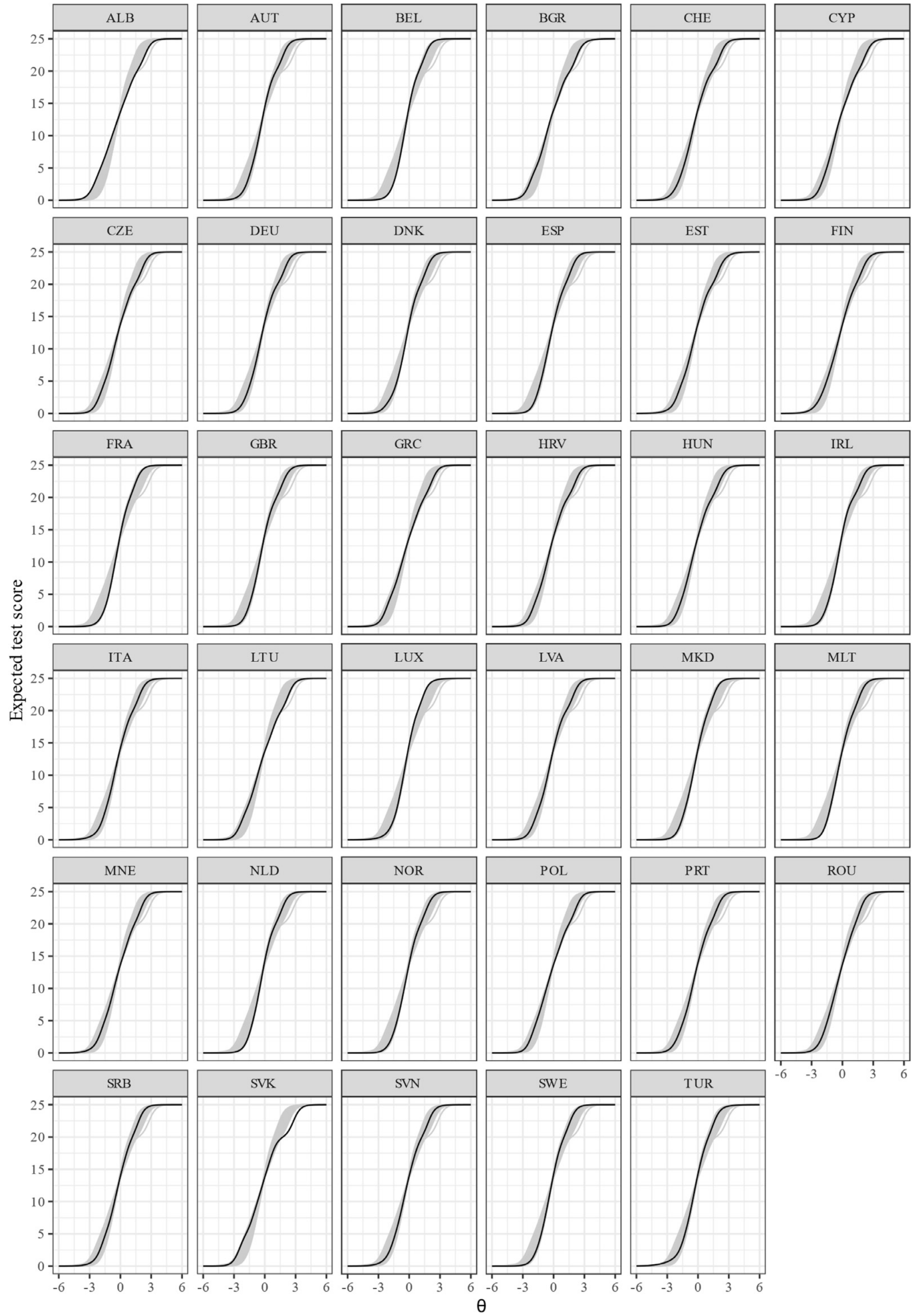

Fig. 5. Test characteristic curves for the GRM after alignment.

In this context, a simple scoring method (e.g., mean or sum scoring) can lead to serious error of inference (e.g., Chen, 2008).[4] Therefore, the manifest norms reported in Topp et al. (2015) have to be regarded with caution and might be misleading. This has also important implications for the cross-national and cross-cultural literature on the WHO-5 that hitherto has not taken into account this liability of the measure and probably suffers from non-valid conclusions (see the various cross-national epidemiology studies in Topp et al., 2015). Thus, future studies should employ a latent variable modeling approach when studying predictors or consequences of the WHO-5.

Model fit to the GRM over countries differed. The *RMSEA* of some countries was relatively high, indicating that the model fitted less well relative to the degrees of freedoms for these countries. Nevertheless,

[4] Sum and mean scores can be regarded as a highly constrained latent variable model with equal loadings (or in IRT terminology: discrimination) parameters across the items (McNeish & Wolf, 2020).

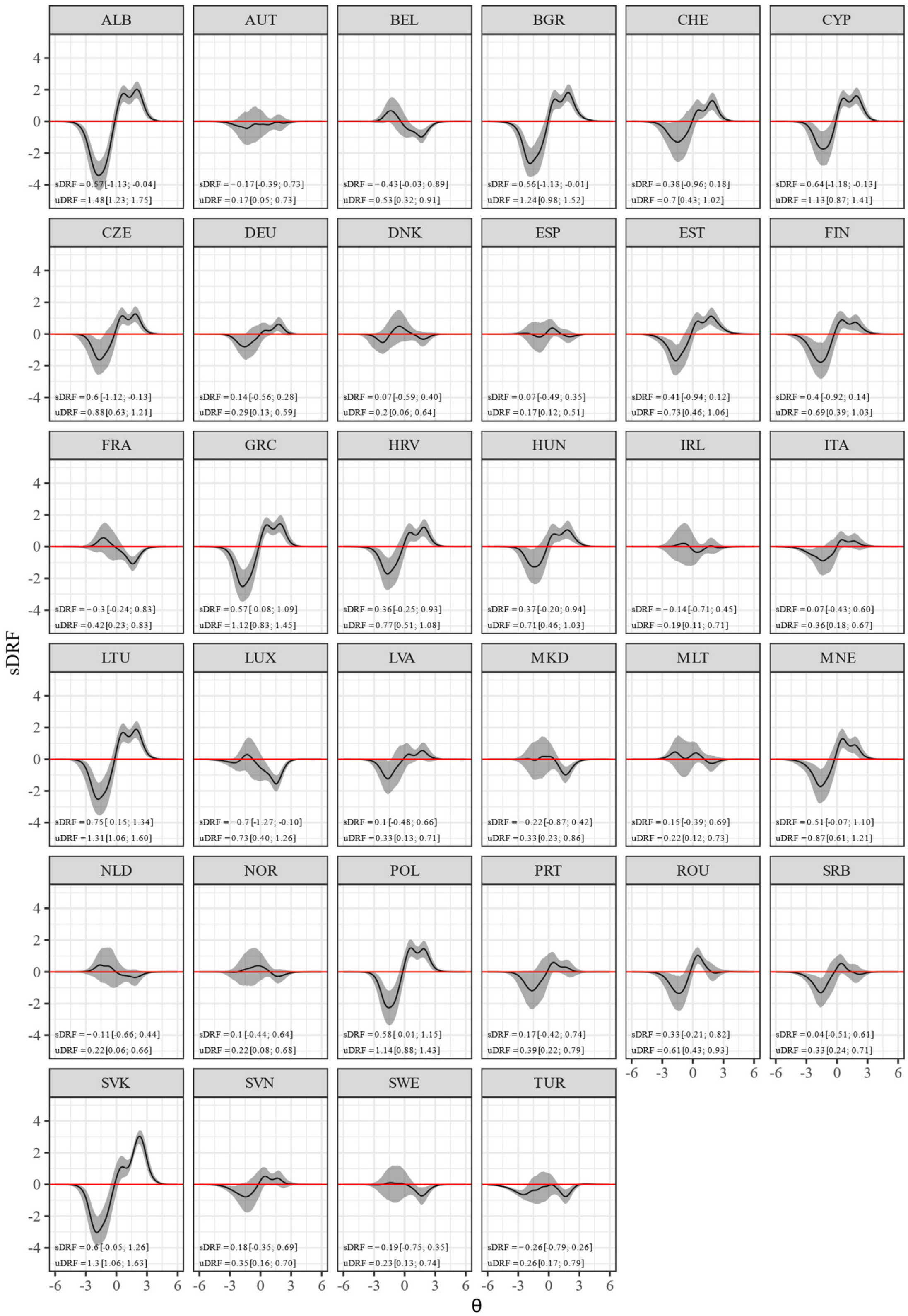

Fig. 6. Differential test functioning for the GRM after alignment (Reference group: GBR).

the analysis of the different assumptions revealed only minor deviation (e.g., local dependency for only a few countries), thus, the IRT assumptions were mostly fulfilled. Test and item information analyses indicated overall (ρ = .83-.93.) as well as at critical points ($\rho_{12.5}$ = .86-.96, ρ_7 = .84-.95) high reliability for all countries. Thus, the present study corroborated the psychometric soundness of the WHO-5 also for countries where an in-depth psychometric analysis was yet missing (e.g., Albania, Czech Republic, Turkey).

Moreover, MI testing and the alignment procedure revealed that there are some non-invariant parameters across countries. Whereas item

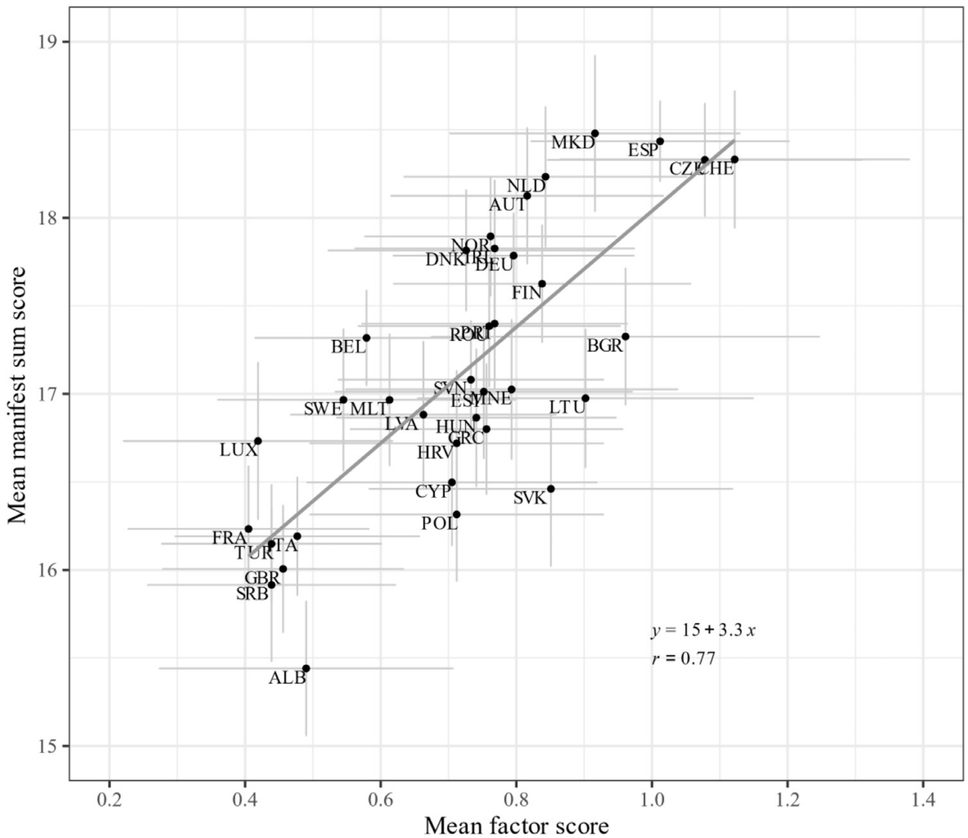

Fig. 7. Scatterplot with means of factor scores and manifest sum scores.

discrimination parameters were approximately invariant, there were some non-invariant thresholds. Especially the first and last threshold of the items showed higher proportions of invariance, indicating less DIF and DTF in the middle range of the latent variable and more on the extreme; a finding that was also corroborated with the in-depth DTF analysis. Manifest sum (or mean) scoring assumes that the measure under study is invariant across groups (Chen, 2008), however, it is important that cross-country studies take possible DTF effects of the WHO-5 into account (for a gentle introduction into measurement invariance testing see e.g., Tay et al., 2015 for an IRT and Pendergast et al., 2017 for a factor analytical approach). For instance, Great Britain and Albania showed an average deviation of 1.48 (99% CI [1.23; 1.75]) points from the WHO-5 expected test scores (ranging between 0 and 25) for the same level on the latent variable. Great Britain yielded expected test scores that were up to 3.40 (99% CI [2.49; 4.33]) points lower for lower levels of the latent variable, whereas expected test scores were up to 2.02 (99% CI [1.45; 2.53]) points higher for higher levels of the latent variable compared with Albania. This means that a respondent from Great Britain with a WHO-5 score of 23.1 (or 2.5) has the same (estimated) level on the latent well-being variable as a respondent from Albania with a WHO-5 score of 21.1 (or 5.9). This exemplifies the problem with simple mean comparisons that becomes particularly relevant when clinical cutoffs are used.

So far, most of the studies that made use of the alignment approach used the assumption of continuous indicators (e.g., Jang et al., 2017). However, modeling the indicators as categorical-ordinal, the alignment method revealed that some thresholds are more invariant than others and that DIF/DTF occurred more at the extremes of the latent variable that are particularly relevant in clinical research. The alignment method allows to compare factor means without satisfying exact scalar invariance that is a hard, if ever, to reach assumption in large-scale cross-cultural research (e.g., Asparouhov and Muthén, 2014; Marsh et al., 2018). Moreover, the alignment procedure has recently been extended to be able to include covariates (Marsh et al., 2018).

4.1. Study strengths, limitations, and outlook

The strength of the study is the large sample size for all included countries (i.e., *n* ranged between 946 and 3,346) and that all samples are nationally representative samples of employed and self-employed individuals. Thus, the analysis should have obtained reasonable item parameter recovery (Ostini et al., 2014). Moreover, the power was large enough to detect even small DIFs between countries (Nguyen et al., 2014). Future studies might consider using samples that also include unemployed persons who are out of jobs or retired. Moreover, future research might seek to test MI for a wider range of countries (e.g., American, Asian, and African countries). In addition, exploring the sources of the non-invariant parameters is an important next step to improve the psychometric properties of the WHO-5 in certain countries (e.g., France).

One limitation of the current study is that the European Working Condition Survey has no external criterion to assess the sensitivity and specificity (e.g., structured clinical interviews) of the WHO-5 to identify depression. However, the validity of the WHO-5 as a depression screening tool has been extensively studied (Krieger et al., 2014; Topp et al., 2015), indicating high sensitivity and specificity across a wide range of research fields and countries. Nevertheless, all research on cutoffs rests on manifestly derived thresholds. Hence, it remains an open empirical question whether latent cutoffs could improve criterion validity.

5. Conclusion

We want to advocate Topp et al.'s (2015) conclusion that the WHO-5 is a psychometrically sound brief measure with non-invasive questions that tap into the subjective well-being of respondents. However, the present study showed that future studies should implement the GRM and that researcher should be aware of measurement non-invariance when they are conducting cross-cultural research. The alignment procedure can help to identify non-invariant parameters across countries.

6. Software information

Most of the data analysis was done in R (Version 4.0.2; R Core Team, 2020). Data transformations were done with the *tidyverse* (Wickham, 2019), *car* (Fox et al., 2020), *combinat* (Chasalow, 2015), *labelled* (Larmarange et al., 2020), and *sjlabelled* (Lüdecke and Ranzolin, 2020) packages. Descriptive statistics were calculated with the *weights* (Pasek and Tahk, 2020) and the *Weighted.Desc.Stat* (Parchami and Bahonar, 2016) packages. Weighted polychoric correlations were calculated with the *wCorr* (Emad and Bailey, 2017) package. Parallel analysis, map test, and description were done with the *psych* (Revelle, 2019) package. Item response analyses were done with the *mirt* (Chalmers, 2012), *nonnest2* (Merkle et al., 2020), and *irtplay* (Lim and Wells, 2020) packages. The graphs were created with the *ggplot2* (Wickham et al., 2020) and *ggpubr* (Kassambara, 2020) packages. The alignment analysis was done in Mplus (v8.0; Muthén and Muthén, 2017) and read in R with the package *MplusAutomation* (Hallquist and Wiley, 2018). Although this study was not preregistered, the R and Mplus scripts used in this article are stored on Open Science Framework (https://osf.io/agfmk/).

Declaration of Competing Interest

None.

Acknowledgements

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jadr.2020.100020.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Allgaier, A.K., Kramer, D., Saravo, B., Mergl, R., Fejtkova, S., Hegerl, U., 2013. Beside the geriatric depression scale: the WHO-five well-being index as a valid screening tool for depression in nursing homes. Int. J. Geriatr. Psychiatry 28, 1197–1204. https://doi.org/10.1002/gps.3944.

Asparouhov, T., Muthén, B., 2014. Multiple-group factor analysis alignment. Struct. Eq. Model. 21, 495–508. https://doi.org/10.1080/10705511.2014.919210.

Bech, P., 1993. Rating scales for psychopathology, health status and quality of life. A compendium on documentation in accordance with the DSM-III-R and WHO systems. Springer, Berlin.

Bech, P., 2012. Clinical Psychometrics. John Wiley and Sons, New York.

Bech, P., Gudex, C., Johansen, K., Staehr, 1996. The WHO (Ten) well-being index: validation in diabetes. Psychother. Psychosom. 65, 183–190. https://doi.org/10.1159/000289073.

Bech, P., Olsen, L.R., Kjoller, M., Rasmussen, N.K., 2003. Measuring well-being rather than the absence of distress symptoms: a comparison of the SF-36 Mental Health subscale and the WHO-Five well-being scale. Int. J. Methods Psychiatr. Res. 12, 85–91. https://doi.org/10.1002/mpr.145.

Boer, D., Hanke, K., He, J., 2018. On detecting systematic measurement error in cross-cultural research: a review and critical reflection on equivalence and invariance tests. J. Cross-Cult. Psychol. 49, 713–734. https://doi.org/10.1177/0022022117749042.

Boye, K., 2009. Relatively different? How do gender differences in well-being depend on paid and unpaid work in Europe? Soc. Indicators Res. 93, 509–525. https://doi.org/10.1007/s11205-008-9434-1.

Brown, A., 2018. Item response theory approaches to test scoring and evaluating the score accuracy,. In: Irwing, P., Hughes, D.J., Booth, T. (Eds.), The Wiley handbook of psychometric testing. A multidisciplinary reference on survey, scale and test development. John Wiley and Sons, New York, pp. 607–638.

Cai, L., Hansen, M., 2013. Limited-information goodness-of-fit testing of hierarchical item factor models. Br. J. Math. Stat. Psychol. 66, 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x.

Cai, L., Monroe, S., 2013. IRT model fit evaluation from theory to practice: Progress and some unanswered questions. Measurement 11, 102–106. https://doi.org/10.1080/15366367.2013.835172.

Cai, L., Monroe, S., 2014. A New Statistic for Evaluating Item Response Theory Models for Ordinal Data. CRESST Report 839. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chalmers, R.P., 2012. mirt: a multidimensional item response theory package for the R environment. J. Statistic. Software 48, 1–29.

Chalmers, R.P., 2018. Model-based measures for detecting and quantifying response bias. Psychometrika 83, 696–732. https://doi.org/10.1007/s11336-018-9626-9.

Chasalow, S., 2015. combinat: combinatorics utilities. In: R package Version 0, pp. 0–8.

Chen, F.F., 2008. What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research.. J. Pers. Soc. Psychol. 95, 1005–1018. https://doi.org/10.1037/a0013193.

Damsbo, A.G., Kraglund, K.L., Buttenschøn, H.N., Johnsen, S.P., Andersen, G., Mortensen, J.K., 2020. Predictors for wellbeing and characteristics of mental health after stroke. J. Affect. Disord. 264, 358–364. https://doi.org/10.1016/j.jad.2019.12.032.

De Ayala, R.J., 2009. The theory and practice of item response theory. Guilford, New York.

DeMars, C.E., 2018. Classical test theory and item response theory. In: Irwing, P., Booth, T, Hughes, D.J. (Eds.), The Wiley Handbook of Psychometric Testing. A Multidisciplinary Reference on Survey, Scale and Test Development. John Wiley and Sons, New York, pp. 49–73.

DeMars, C.E., 2020. Alignment as an alternative to anchor purification in DIF analyses. Struct. Eq. Model. 27, 56–72. https://doi.org/10.1080/10705511.2019.1617151.

Edwards, M.C., Houts, C.R., Cai, L., 2018. A diagnostic procedure to detect departures from local independence in item response theory models. Psychol. Methods 23, 138–149. http://dx.doi.org/10.1037/met0000121.

Edwards, M.C., Wirth, R.J., Houts, C.R., Xi, N., 2012. Categorical data in the structural equation modeling framework. In: Hoyle, R.H. (Ed.), Handbook of Structural Equation Modeling. Guilford, New York, pp. 195–208.

El-Den, S., Chen, T.F., Gan, Y.L., Wong, E., O'Reilly, C.L., 2018. The psychometric properties of depression screening tools in primary healthcare settings: A systematic review. J. Affect. Disord. 225, 503–522. https://doi.org/10.1016/j.jad.2017.08.060.

Elholm, B., Larsen, K., Hornnes, N., Zierau, F., Becker, U., 2011. Alcohol withdrawal syndrome: symptom-triggered versus fixed-schedule treatment in an outpatient setting. Alcohol Alcohol. 46, 318–323. https://doi.org/10.1093/alcalc/agr020.

Emad, A., Bailey, P., 2017. wCorr: weighted correlations. R package Version 1.9.1. https://cran.r-project.org/web/packages/wCorr/index.html.

Eurofound (Ed.)., 2015a. 6th European working conditions survey – technical report. Working Paper. http://www.eurofound.europa.eu/sites/default/files/ef_survey/field_ef_documents/6th_ewcs_-_technical_report.pdf.

Eurofound (Ed.)., 2015b. 6th European working conditions survey – weighting report. Working Paper https://www.eurofound.europa.eu/sites/default/files/ef_survey/field_ef_documents/6th_ewcs_2015_-_weighting_report.pdf.

Eurofound (Ed.)., 2017. European quality of life survey 2016: quality of life, quality of public services, and quality of society. Publications Office of the European Union, Luxembourg.

Eurofound (Ed.)., 2020. Living, working and COVID-19: first findings. April 2020. European Foundation for the Improvement of Living and Working Conditions. https://www.eurofound.europa.eu/topic/covid-19.

European Foundation for the Improvement of Living and Working Conditions. European working conditions survey, 2015. [data collection] Data Service. SN: 8098.

Finch, W.H., 2016. Detection of differential item functioning for more than two groups: a Monte Carlo comparison of methods. Appl. Measur. Educ. 29, 30–45. http://doi.org/10.1080/08957347.2015.1102916.

Fowler, J.C., Clapp, J.D., Madan, A., Allen, J.G., Frueh, B.C., Fonagy, P., Oldham, J.M., 2018. A naturalistic longitudinal study of extended inpatient treatment for adults with borderline personality disorder: An examination of treatment response, remission and deterioration. J. Affect. Disord. 235, 323–331. https://doi.org/10.1016/j.jad.2017.12.054.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G.Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nillson, H., R-Core., 2020. Car: companion to applied regression. In: R Package Version 3, pp. 0–8.

Garrido, L.E., Abad, F.J., Ponsoda, V., 2011. Performance of Velicer's minimum average partial factor retention method with categorical variables. Educ. Psychol. Measur. 71, 551–570. https://doi.org/10.1177/0013164410389489.

Garrido, L.E., Abad, F.J., Ponsoda, V., 2013. A new look at Horn's parallel analysis with ordinal variables. Psychol. Methods 18, 454–474. https://doi.org/10.1037/a0030005.

Halliday, J.A., Hendrieckx, C., Busija, L., Browne, J.L., Nefs, G., Pouwer, F., Speight, J., 2017. Validation of the WHO-5 as a first-step screening instrument for depression in adults with diabetes: results from diabetes MILES–Australia. Diabetes Res. Clin. Pract. 132, 27–35. https://doi.org/10.1016/j.diabres.2017.07.005.

Hallquist, M.N., Wiley, J.F., 2018. Mplusautomation: an R package for facilitating large-scale latent variable analyses in Mplus. Struct. Eq. Model. 25, 621–638. https://doi.org/10.1080/10705511.2017.1402334.

Harkness, J.A., Van de Vijver, F.J.R., Mohler, P.P. (Eds.), 2003. Cross-Cultural Survey Methods. John Wiley and Sons, New York.

Hartwig, E.M., Rufino, K.A., Palmer, C.A., Shepard, C., Alfano, C.A., Schanzer, B., Mathew, S.J., Patriquin, M.A., 2019. Trajectories of self-reported sleep disturbance across inpatient psychiatric treatment predict clinical outcome in comorbid major depressive disorder and generalized anxiety disorder. J. Affect. Disord. 251, 248–255. https://doi.org/10.1016/j.jad.2019.03.069.

Horn, J.L., 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30, 179–185. https://doi.org/10.1007/BF02289447.

Huelmann, T., Debelak, R., Strobl, C., 2020. A comparison of aggregation rules for selecting anchor items in multigroup DIF analysis. J. Educ. Measur. 57, 185–215. https://doi.org/10.1111/jedm.12246.

Jang, S., Kim, E.S., Cao, C., Allen, T.D., Cooper, C.L., Lapierre, L.M., Abarca, N., 2017. Measurement invariance of the satisfaction with life scale across 26 countries. J. Cross-Cult. Psychol. 48, 560–576. https://doi.org/10.1177/0022022117697844.

Kang, T., Chen, T.T., 2011. Performance of the generalized SX 2 item fit index for the graded response model. Asia Pacific Educ. Rev. 12, 89–96. https://doi.org/10.1007/s12564-010-9082-4.

Kang, T., Cohen, A.S., Sung, H.J., 2009. Model selection indices for polytomous items. Appl. Psychol. Measur. 33, 499–518. https://doi.org/10.1177/0146621608327800.

Kassambara, A., 2020. ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0. https://cran.r-project.org/web/packages/ggpubr/index.html.

Killikelly, C., Lorenz, L., Bauer, S., Mahat-Shamir, M., Ben-Ezra, M., Maercker, A., 2019. Prolonged grief disorder: Its co-occurrence with adjustment disorder and post-traumatic stress disorder in a bereaved Israeli general-population sample. J. Affect. Disord. 249, 307–314. https://doi.org/10.1016/j.jad.2019.02.014.

Kim, E.S., Cao, C., Wang, Y., Nguyen, D.T., 2017. Measurement invariance testing with many groups: a comparison of five approaches. Struct. Eq. Model. 24, 524–544. doi:10.1080/10705511.2017.1304822.

Kline, R.B., 2015. Principles and Practice of Structural Equation Modeling, 4th ed. Guilford, New York.

Krieger, T., Zimmermann, J., Huffziger, S., Ubl, B., Diener, C., Kuehner, C., Holtforth, M, Grosse, 2014. Measuring depression with a well-being index: further evidence for the validity of the WHO Well-Being Index (WHO-5) as a measure of the severity of depression. J. Affect. Disord. 156, 240–244. doi:10.1016/j.jad.2013.12.015.

Larmarange, J., Ludecke, D., Wickham, H., Bojanowski, M., Briatte, F., 2020. labelled: manipulating labelled data. R package Version 2.5.0. https://cran.r-project.org/web/packages/labelled/index.html.

Lim, H., Wells, C.S., 2020. irtplay: unidimensional item response theory modeling. R package Version 1.6.1. https://cran.r-project.org/web/packages/irtplay/index.html.

Löwe, B., Spitzer, R.L., Gräfe, K., Kroenke, K., Quenter, A., Zipfel, S., Buchholz, C., Witte, S., Herzog, W., 2004. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J. Affect. Disord. 78, 131–140. https://doi.org/10.1016/S0165-0327(02)00237-9.

Lucas-Carrasco, R., Allerup, P., and Bech, P. (2012). The validity of the WHO-5 as an early screening for apathy in an elderly population. Current gerontology and geriatrics research, 2012. https://doi.org/10.1155/2012/171857.

Lüdecke, D., Ranzolin, D., 2020. sjlabelled: labelled data utility functions. R package Version 1.1.6. https://cran.r-project.org/web/packages/sjlabelled/index.html.

Marsh, H.W., Guo, J., Parker, P.D., Nagengast, B., Asparouhov, T., Muthén, B., 2018. What to do when scalar invariance fails: the extended alignment method for multi-group factor analysis comparison of latent means across many groups. Psychol. Methods 23, 524–545. doi:10.1037/met0000113.

Masters, G.N., 1982. A Rasch model for partial credit scoring. Psychometrika 47, 149–174. https://doi.org/10.1007/bf02296272.

Maydeu-Olivares, A., Joe, H., 2014. Assessing approximate fit in categorical data analysis. Multivariate Behav. Res. 49, 305–328. https://doi.org/10.1080/00273171.2014.911075.

McDowell, I., 2010. Measures of self-perceived well-being. J. Psychosom. Res. 69, 69–79. https://doi.org/10.1016/j.jpsychores.2009.07.002.

McNeish, D., Wolf, M.G., 2020. Thinking twice about sum scores. Behav. Res. Methods. https://doi.org/10.3758/s13428-020-01398-0. in press.

Merkle, E., Schneider, L., Bae, S., 2020. nonnest2: Tests of non-nested models. In: R package Version 0, p. 5–5, https://cran.r-project.org/web/packages/nonnest2/index.html.

Millsap, R.E., 2011. Statistical approaches to measurement invariance. Routledge, New York.

Monroe, S., Cai, L., 2015. Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. Multivariate Behav. Res. 50, 569–583. https://doi.org/10.1080/00273171.2015.1032398.

Muraki, E., 1992. A generalized partial credit model: application of an EM algorithm. Appl. Psychol. Measur. 16, 159–176. https://doi.org/10.1177/014662169201600206.

Muthén, B., Asparouhov, T., 2014. IRT studies of many groups: the alignment method. Front. Psychol. 5, 978. https://doi.org/10.3389/fpsyg.2014.00978.

Muthén, L.K., Muthén, B.O., 2017. Mplus user's guide, 8th Ed.. Los Angeles, CA, Muthén and Muthén.

Nguyen, T.H., Han, H.R., Kim, M.T., Chan, K.S., 2014. An introduction to item response theory for patient-reported outcome measurement. Patient 7, 23–35. https://doi.org/10.1007/s40271-013-0041-0.

Nicolucci, A., Kovacs Burns, K., Holt, R.I., Comaschi, M., Hermanns, N., Ishii, H., Kokoszka, A., Pouwer, F., Skovlund, S.E., Stuckey, H., Tarkun, I., Vallis, M., Wens, J., Peyrot, M., 2013. Diabetes attitudes, wishes and needs second study (DAWN2™): cross-national benchmarking of diabetes-related psychosocial outcomes for people with diabetes. Diabet. Med. 30, 767–777. https://doi.org/10.1111/dme.12245.

Ostini, R., Finkelman, M., Nering, M., 2014. Selecting among polytomous IRT models. In: Reise, S.P., Revicki, D.A. (Eds.), Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment. Routledge, New York, pp. 285–304.

Parchami, A., Bahonar, S., 2016. Weighted.Desc.Stat: weighted descriptive statistics. R package Version 1.0. https://cran.r-project.org/web/packages/Weighted.Desc.Stat/index.html.

Pasek, J., Tahk, A., Culter, G., Schwemmle, M., 2020. weights: weighting and weighted statistics. R package Version 1.0.1. https://cran.r-project.org/web/packages/weights/index.html.

Pendergast, L.L., von der Embse, N., Kilgus, S.P., Eklund, K.R., 2017. Measurement equivalence: a non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. J. School Psychol. 60, 65–82. https://doi.org/10.1016/j.jsp.2016.11.002.

R Core Team, 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Reise, S.P., Widaman, K.F., Pugh, R.H., 1993. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol. Bull. 114, 552–566. https://doi.org/10.1037/0033-2909.114.3.552.

Revelle, W., 2019. psych: Procedures for psychological, psychometric, and personality research. R package Version 1.9. 12.31. https://cran.r-project.org/web/packages/psych/index.html.

Rose, T., Joe, S., Williams, A., Harris, R., Betz, G., Stewart-Brown, S., 2017. Measuring mental wellbeing among adolescents: a systematic review of instruments. J. Child Family Stud. 26, 2349–2362. https://doi.org/10.1007/s10826-017-0754-0.

Saipanish, R., Lotrakul, M., Sumrithe, S., 2009. Reliability and validity of the Thai version of the WHO-five well-being index in primary care patients. Psychiatry Clin. Neurosci. 63, 141–146. https://doi.org/10.1111/j.1440-1819.2009.01933.x.

Samejima, F., 1969. Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement No. 17.

Schneider, L., Chalmers, R.P., Debelak, R., Merkle, E.C., 2019. Model selection of nested and non-nested item response models using Vuong tests. Multivariate Behav. Res.. https://doi.org/10.1080/00273171.2019.1664280. (in press).

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statis. 6, 461–464. https://doi.org/10.1214/aos/1176344136.

Sijtsma, K., 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74, 107–120. https://doi.org/10.1007/s11336-008-9101-0.

Sisask, M., Värnik, A., Kolves, K., Konstabel, K., Wasserman, D., 2008. Subjective psychological well-being (WHO-5) in assessment of the severity of suicide attempt. Nord. J. Psychiatry 62, 431–435. https://doi.org/10.1080/08039480801959273.

Sischka, P.E., Schmidt, A.F., Steffgen, G., 2020. Further evidence for criterion validity and measurement invariance of the Luxembourg Workplace Mobbing Scale. Eur. J. Psychol. Ass. 36, 32–43. https://doi.org/10.1027/1015-5759/a000483.

Steenkamp, J.B.E., Baumgartner, H., 1998. Assessing measurement invariance in cross-national consumer research. J. Consum. Res. 25, 78–90. https://doi.org/10.1086/209528.

Tay, L., Meade, A.W., Cao, M., 2015. An overview and practical guide to IRT measurement equivalence analysis. Org. Res. Methods 18, 3–46. https://doi.org/10.1177/1094428114553062.

Topp, C.W., Østergaard, S.D., Søndergaard, S., Bech, P., 2015. The WHO-5 well-being index: a systematic review of the literature. Psychother. Psychosom. 84, 167–176. doi:10.1159/000376585.

Van Gestel, Y.R.B.M., Voogd, A.C., Vingerhoets, A.J.J.M., Mols, F., Nieuwenhuijzen, G.A.P., van Driel, O.R., van Berlo, C.L.H., van de Poll-Franse, L.V., 2007. A comparison of quality of life, disease impact and risk perception in women with invasive breast cancer and ductal carcinoma in situ. Eur. J. Cancer 43, 549–556. https://doi.org/10.1016/j.ejca.2006.10.010.

Velicer, W.F., 1976. Determining the number of components from the matrix of partial correlations. Psychometrika 41, 321–327. https://doi.org/10.1007/BF02293557.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333. https://doi.org/10.2307/1912557.

Wells, C.S., Hambleton, R.K., 2016. Model fit with residual analyses. In: van der Linden, W.J (Ed.), Handbook of item response theory, volume two: Statistical tools. CRC Press, Boca Raton, pp. 395–413.

Whittaker, T.A., 2013. The impact of noninvariant intercepts in latent means models. Struct. Eq. Model. 20, 108–130. doi:10.1080/10705511.2013.742397.

Wickham, H. 2019., tidyverse: easily install and load the 'Tidyverse'. R package version 1.3.0. https://cran.r-project.org/web/packages/tidyverse/index.html.

Wickham H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., RStudio, 2020. ggplot2: create elegant data visualisations using the grammar of graphics. R package version 3.3.2. https://cran.r-project.org/web/packages/ggplot2/index.html.

Wirth, R.J., Edwards, M.C., 2007. Item factor analysis: current approaches and future directions. Psychol. Methods 12, 58–79. https://doi.org/10.1037/1082-989X.12.1.58.

World Health Organization, 1998. Well-Being measures in primary health care: the DepCare Project. WHO Regional Office for Europe, Copenhagen.

World Health Organization, 2020. Basic documents, 49th ed. World Health Organization, Geneva.