P. E. Sischka

University of Luxembourg
Contact: Philipp Sischka, University of Luxembourg, Department of Behavioural and Cognitive Sciences, Porte des Sciences, L-4366 Esch-sur-Alzette, philipp.sischka@uni.lu

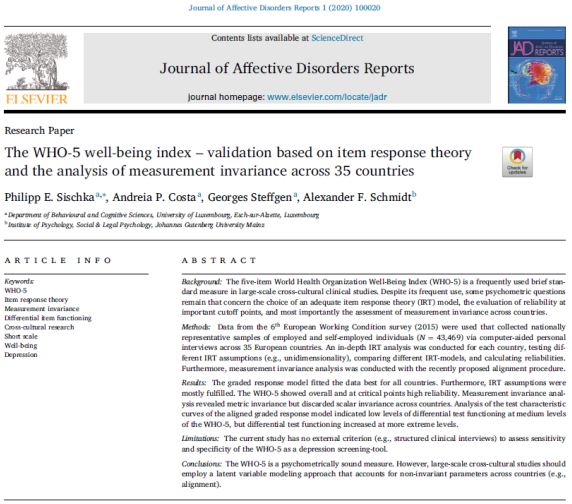
The cross-country validation of the WHO-5 well-being index with item response theory and the alignment procedure

RC33 Logic and Methodology in Sociology

RC33 Online conference 2021

Session 7: Assessing the Quality of Survey Data I

Trier, 2021, September 08th



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).

State of the art

The WHO-5 well-being index

- One of the most widely used measure for subjective well-being (Topp et al., 2015)
  - 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)
  - occupational psychology (Sischka et al., 2020)
  - …
- Measures a global hedonic dimension of well-being (Bech, 2012)

State of the art

Psychometric properties of the WHO-5

- Research on (~213 studies; Topp et al., 2015)
  - sensitivity and specificity to detect depression ($M_{sensitivity}$ = .86, $M_{specificity}$ = .81)
  - internal consistency (Cronbach's Alpha)
  - unidimensionality (EFA, PCA, CFA)
  - single IRT models (Mokken scaling, partial credit model, graded response model)
- Lack of research and study aim of current study
  - adequate IRT model (partial credit model, generalized partial credit model, graded response model)
  - reliability (at important cutoffs)
  - measurement invariance (across countries)

Method

Survey design and participants

- Survey design
  - Data from the European Working Condition Survey 2015
  - assessment of working conditions of employees and self-employed across Europe (35 countries) within nationally representative samples
  - survey conducted via CAPI
  - multi-stage sample selection process (complex survey sampling)
- Participants
  - 43,469 employees and self-employed (946-3346 respondents per country)
  - 49.6% females, *n* = 21,553
  - Age: 15 to 89 years (*M* = 43.3, *SD* = 12.7)

Method

Measure: WHO-5 well-being index

Instructions: Please indicate for each of the 5 statements which is closest to how you have been feeling over the past 2 weeks.							
	Over the past 2 weeks…	At no time	Some of the time	Less than half of the time	More than half of the time	Most of the time	All of the time
1	... I have felt cheerful and in good spirits.	0	1	2	3	4	5
2	... I have felt calm and relaxed.	0	1	2	3	4	5
3	... I have felt active and vigorous.	0	1	2	3	4	5
4	... I woke up feeling fresh and rested.	0	1	2	3	4	5
5	... my daily life has been filled with things that interest me.	0	1	2	3	4	5

Scaling procedure: summing up the five items. Theoretical range between 0 (absence of well-being) to 25 (maximal well-being)

Method

Statistical analyses

- Comparing different IRT models (PCM, GPCM, GRM):
  - AIC, BIC (+ sample-size adjusted versions)
  - Vuong Test (Vuong, 1989, Schneider et al., 2019)
  - Change in R^2
  - Fit indices (C_2 test statistic with corresponding fit indices; Cai & Monroe, 2014)
- Testing IRT assumptions / psychometric properties
  - Unidimensionality (parallel analysis and minimum average partial method)
  - Local indecency (Jackknife Slope Index; Edwards et al., 2018)
  - Monotonicity (Raw residual plots)
  - Item fit (Generalized S-X2 item fit index (Kang and Chen, 2011) and RMSEA)
  - Item/test characteristic curves and Item/test information functions

Method

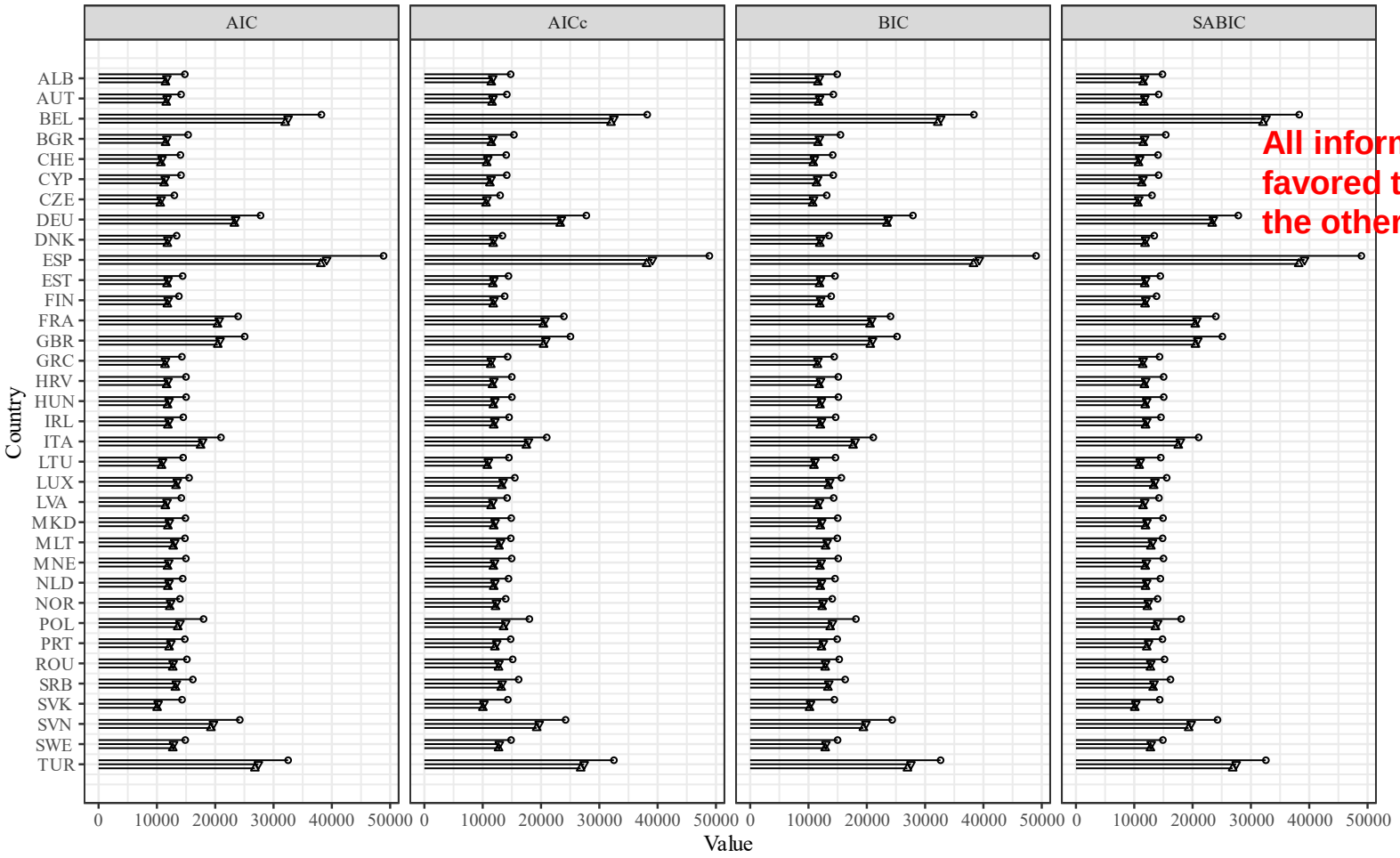
Statistical analyses

- Measurement invariance testing
  - Multigroup IRT analysis (configural, metric, scalar invariance model)
  - Alignment procedure (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014)
    - The alignment method “serves the joint purposes of scale linking and purification, without literally deleting items from the linking” (DeMars, 2020, p. 56)
    - Identification of invariant and non-invariant parameters
    - Global measure of (non-)invariance: R^2 measure (0 and 1)
  - Differential response functioning (DRF) statistics (Chalmers, 2018) as effect size measure for differential test functioning

Results

Comparing PCM and GPCM: ΔR^2 between .112 and .288.
Comparing GPCM and GRM: ΔR^2 between .005 and .027.

Model comparison for the PCM, GPCM, and GRM.



All information criteria favored the GRM over the other models.

Results

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

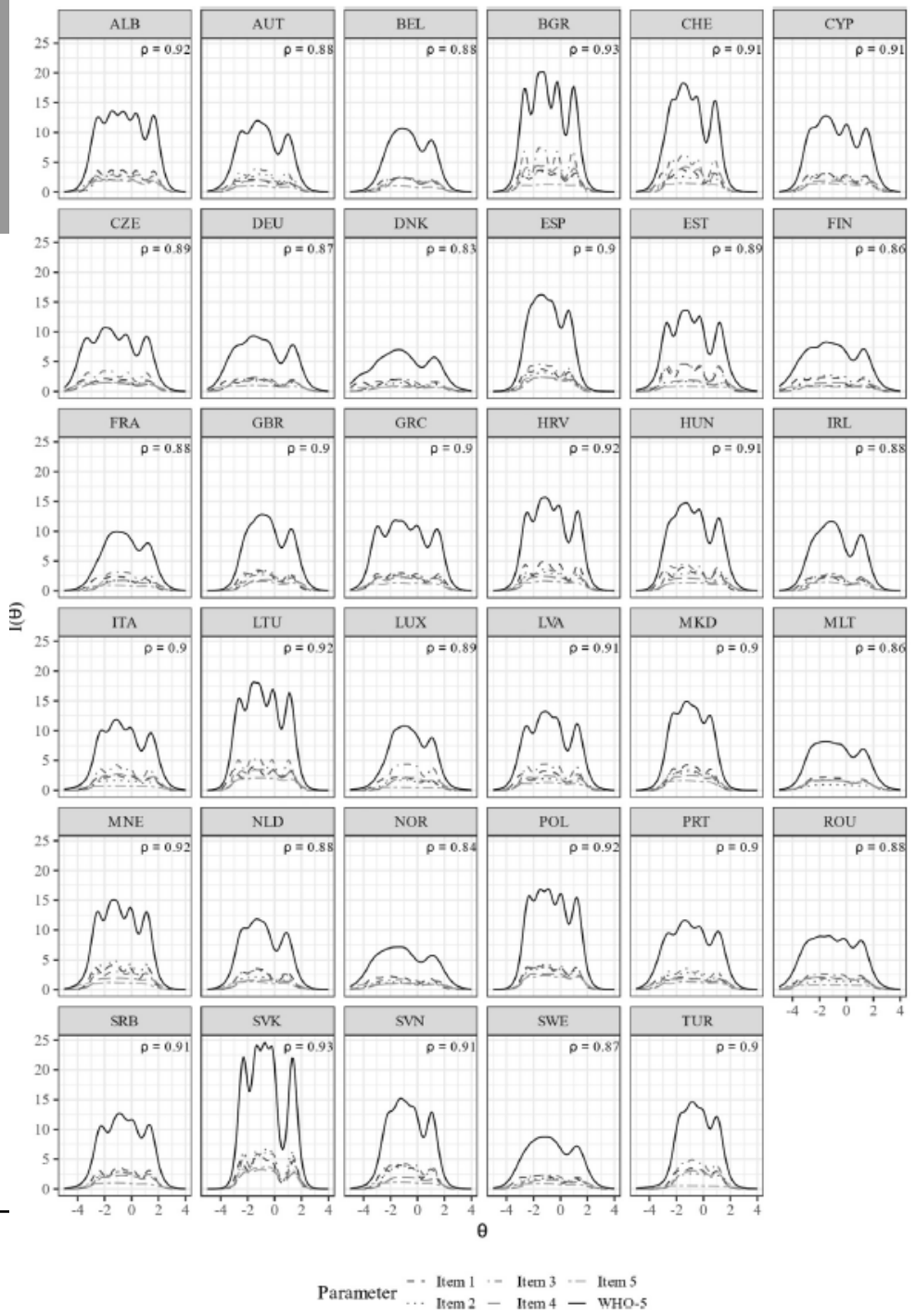
Notes. $df = 5$; *RMSEA* = root mean squared error of approximation; *SRMR* = standardized root mean square residual; *TLI* = Tucker-Lewis index; *CFI* = comparative fit index.

With the execption of some RMSEA values, all fit indices showed adequate fit of the GRM model.

Results

Item and test information functions for the GRM.

$$reliability = 1 - \frac{1}{test\ information}$$



Notes. ρ = Empirical marginal reliability.

Results

Multigroup IRT analysis.

Form of invariance	C_2	*p*	*df*	*RMSEA*	*SRMR*	*CFI*	*TLI*
Configural invariance	2,073.244	0.000	175	.016	.019-.074	.988	.977
Metric invariance	3,248.038	0.000	311	.015	.026-.131	.982	.980
Scalar invariance	17,738.548	0.000	1127	.018	.026-.188	.898	.968
Δ Configural – metric				-.001		-.006	+.003
Δ Metric – scalar				+.002		-.090	-.009

Item discrimination parameters are nearly invariant (metric invariance was confirmed).
Item thresholds parameters are not invariant (scalar invariance had to be rejected).

Notes. *MLR* estimator.

Results

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%)

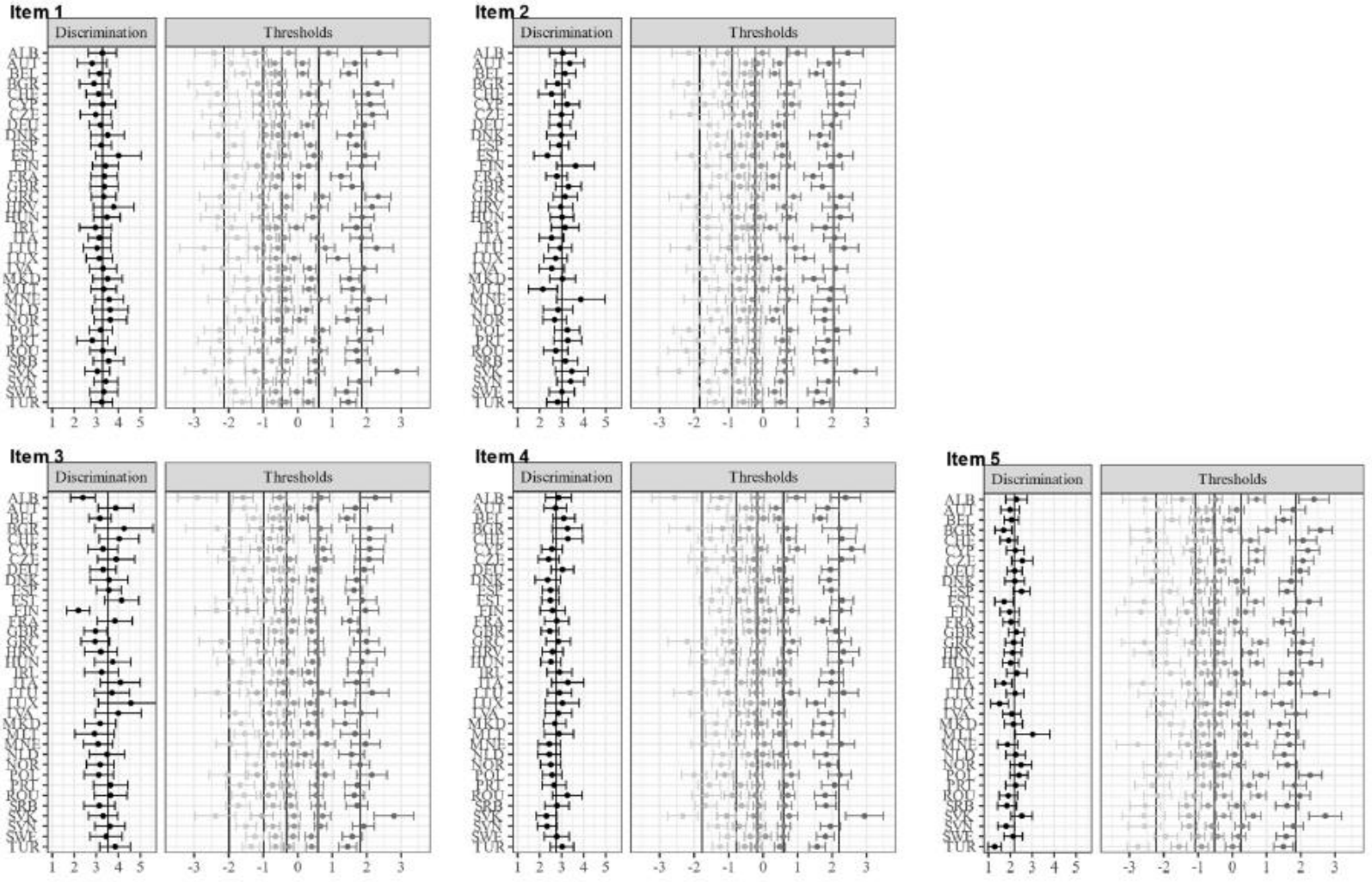
Notes. *MLR* estimator; FREE approach.

Item discrimination parameters nearly invariant.

Item threshold parameters for the lower and upper categories showed higher amount of non-invariance.

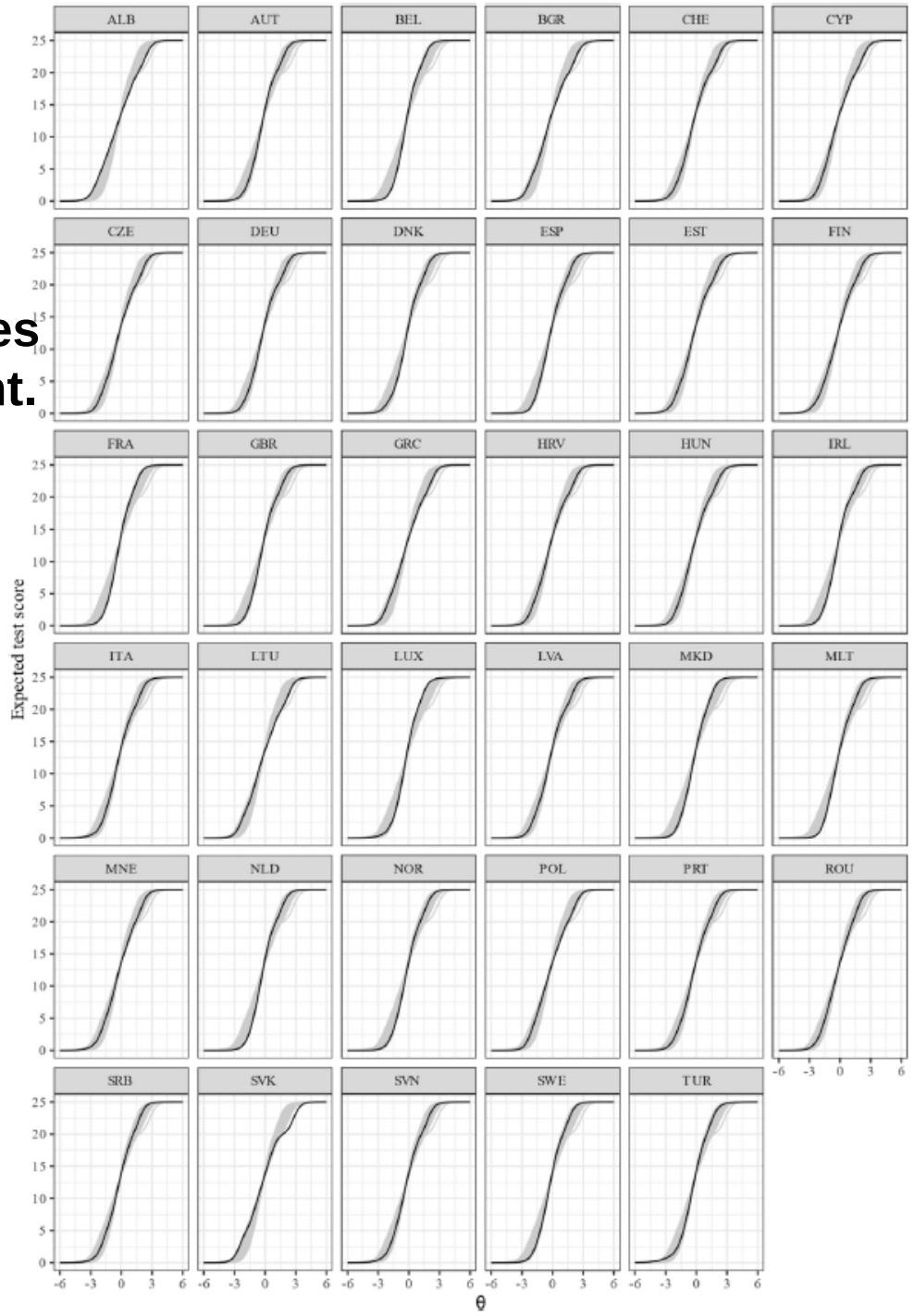
Results

Item parameter for the GRM after alignment.



Results

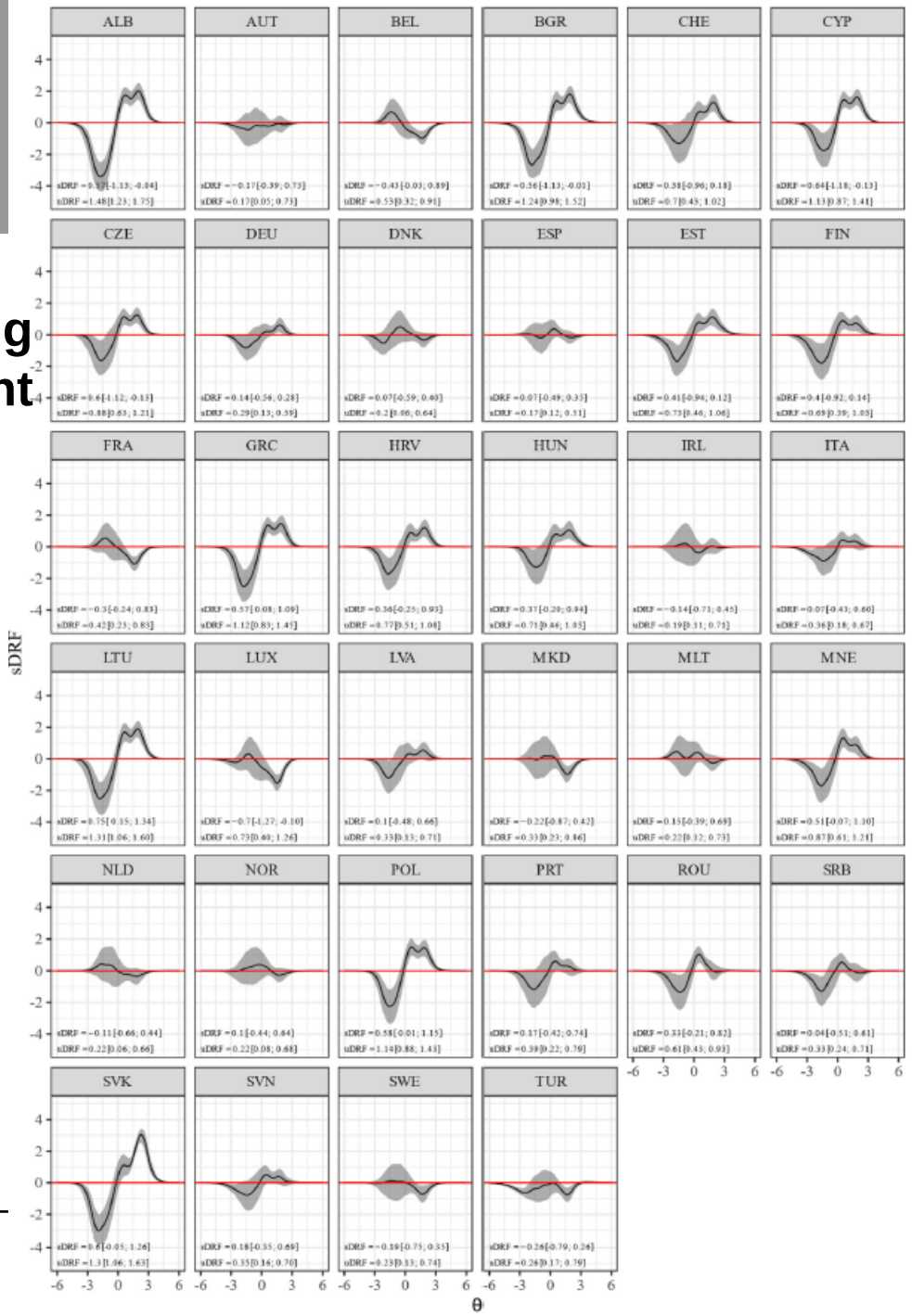
Test characteristic curves for the GRM after alignment.



Results

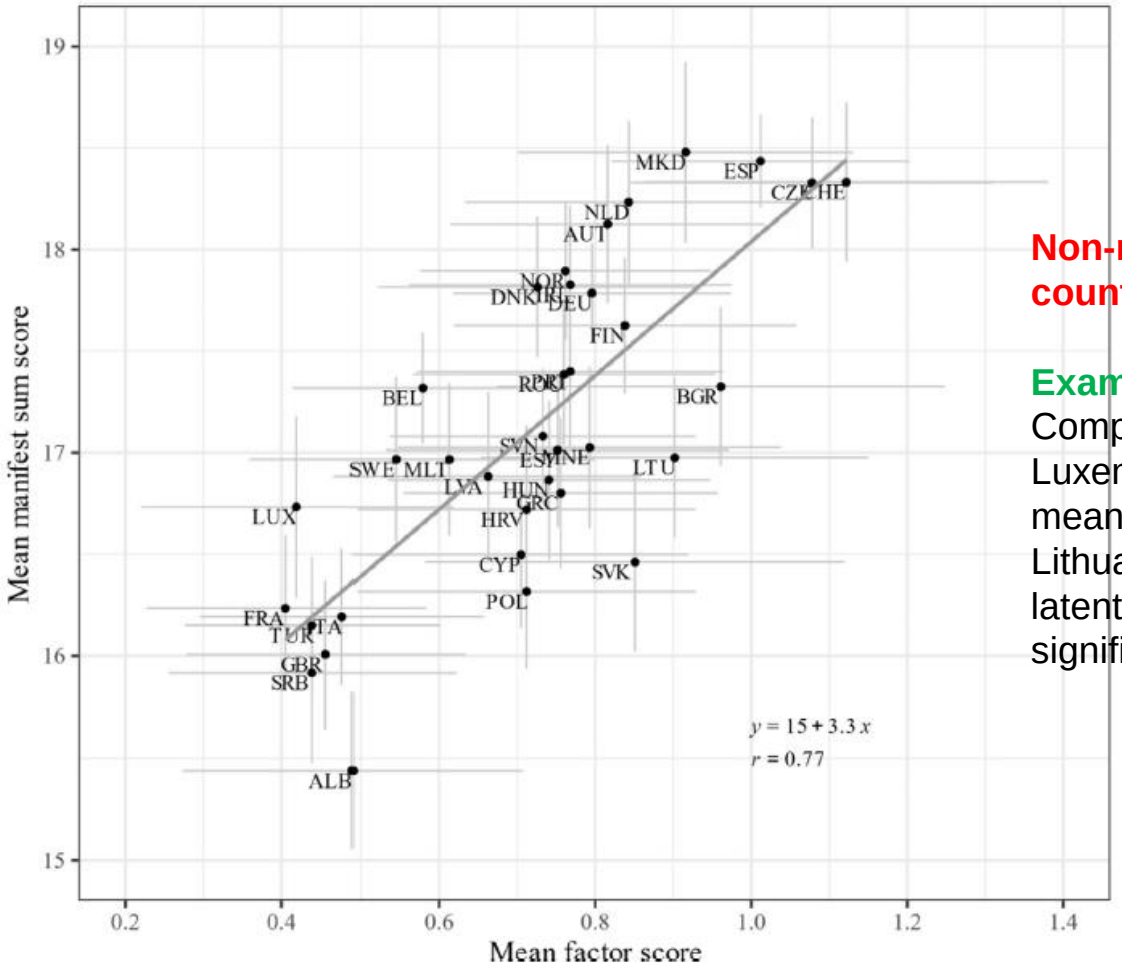
Differential test functioning for the GRM after alignment (Reference group: GBR).

Reading example:
Respondents from Great Britain and Albania who have the same level on the latent well-being variable will have quite different WHO-5 sum scores (2.5 vs. 5.9 at the lower level and 23.1 vs. 21.1 at the upper level).



Results

Scatterplot with means of factor scores and manifest sum scores.



Non-negligible change in the country rank order.

Example

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.

Discussion

Summary

- Every criterion favored the GRM over the other IRT models.
- IRT assumptions (unidimensionality, local independence, monotonicity, item fit) could be confirmed.
- Test information analyses indicated overall as well as at critical points high reliability for all countries.
- Measurement invariance testing confirmed configural and metric invariance but discarded scalar invariance.
- The alignment procedure and the DRF statistics revealed that differential test functioning occurred more at the extreme.

Discussion

Study strengths, limitations, and outlook

- Strength: Large sample size (for all included countries)
- Limitation: no external criterion to assess sensitivity and specificity of the WHO-5 to identify depression
- Outlook: Testing the WHO-5 in unemployed persons and across a wider range of countries (e.g., African)

Conclusion

- WHO-5 is a psychometrically sound brief measure of subjective well-being.
- Cross-cultural research should employ a latent variable approach and consider non-invariant parameters across countries.

Thank you for your attention!

Any questions?

Email: philipp.sischka@uni.lu

(R and Mplus scripts are stored on Open Science Framework https://osf.io/agfmk/).

References

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

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

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.

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

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.

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

References

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.

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.

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) .

Sisask, M., Värnik, A., Kolves, K., Konstabel, K., Wasserman, D., 2008. Subjective psycho- logical 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.

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.

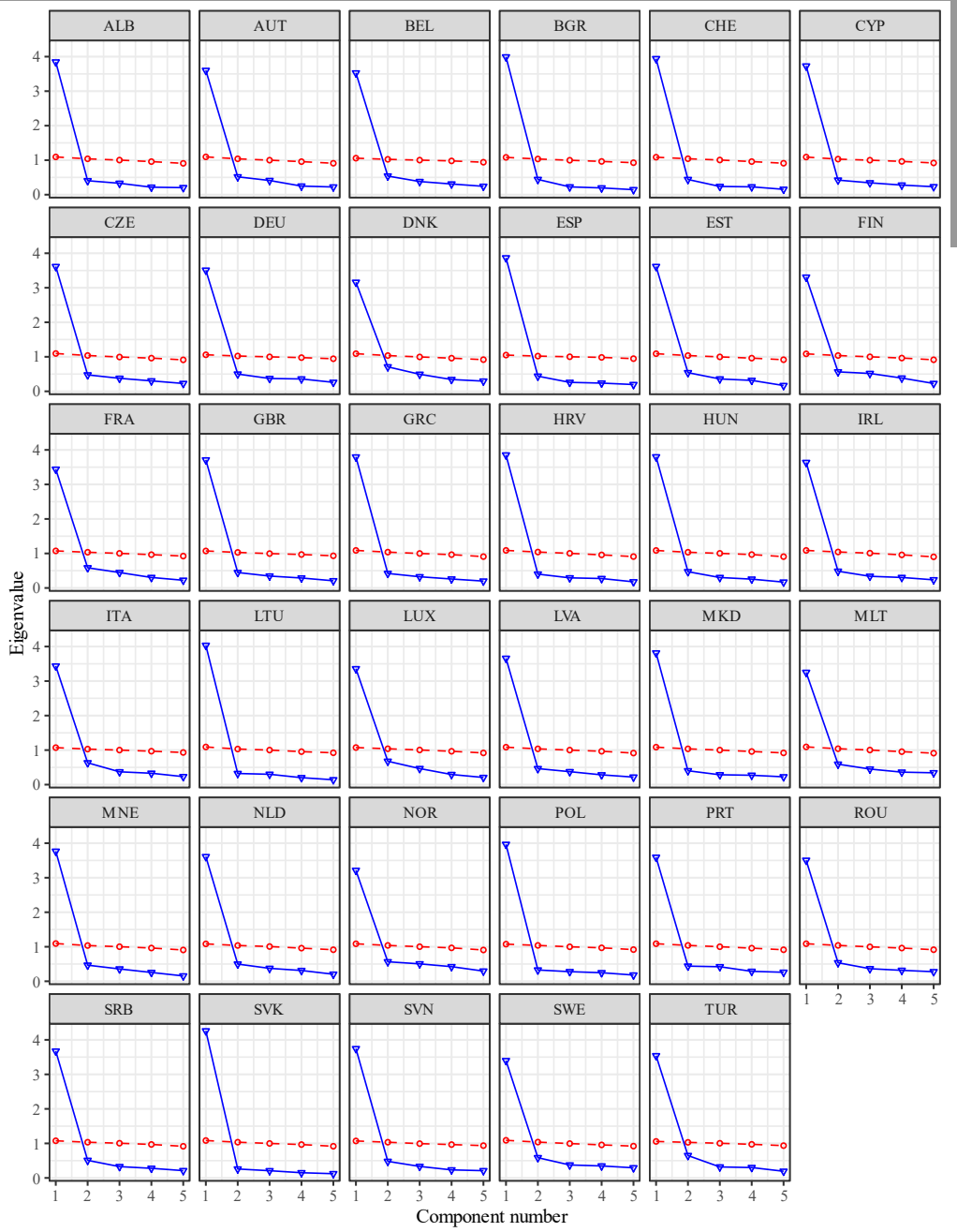
Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333. https://doi.org/10.2307/1912557.

Appendix

Appendix

Parallel analysis.

Unidimensionality is confirmed for all countries.

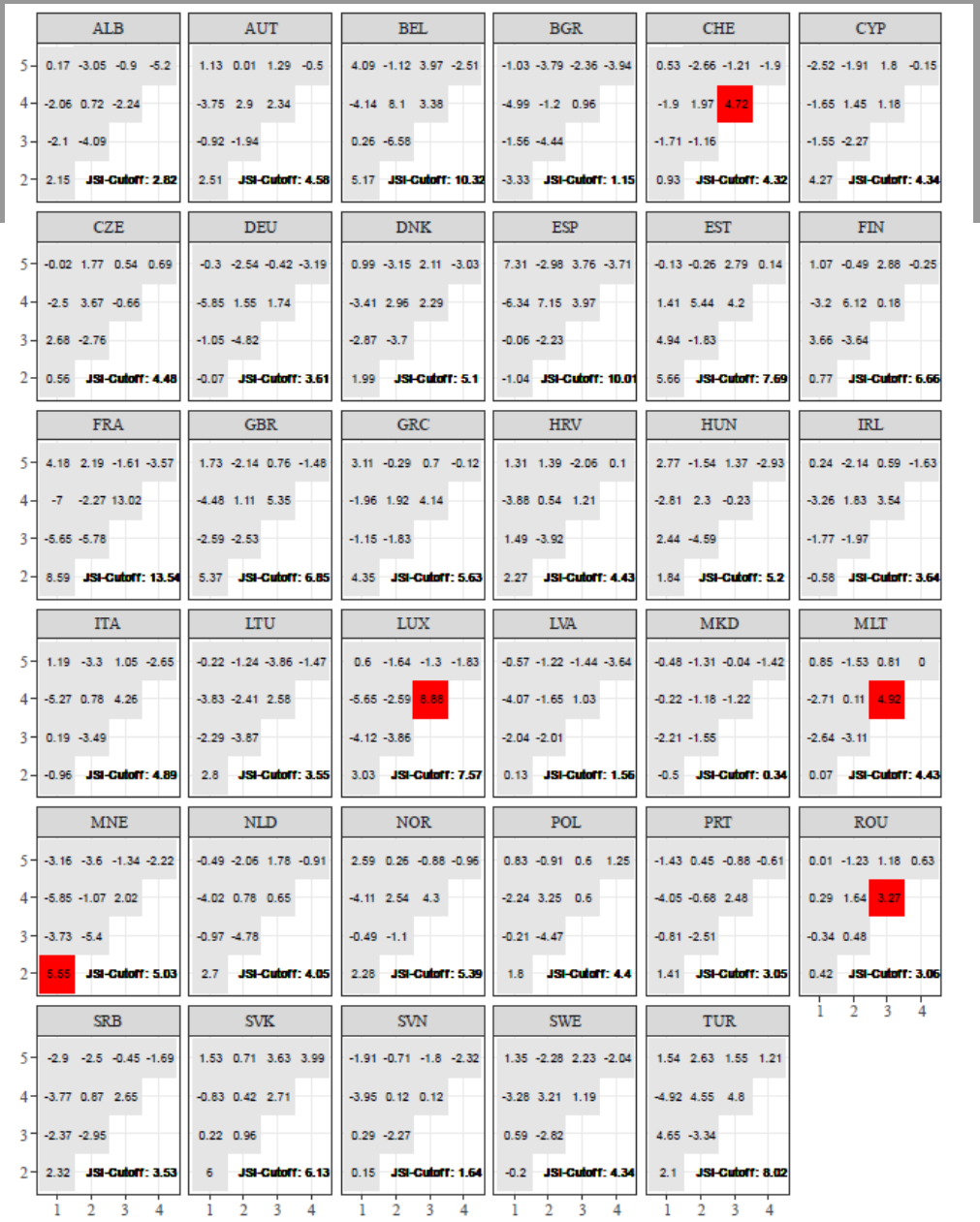


Notes. Blue line: PC actual data; Red line: PC resampled data.

Appendix

Test for local dependency.

Items are mostly locally independent.

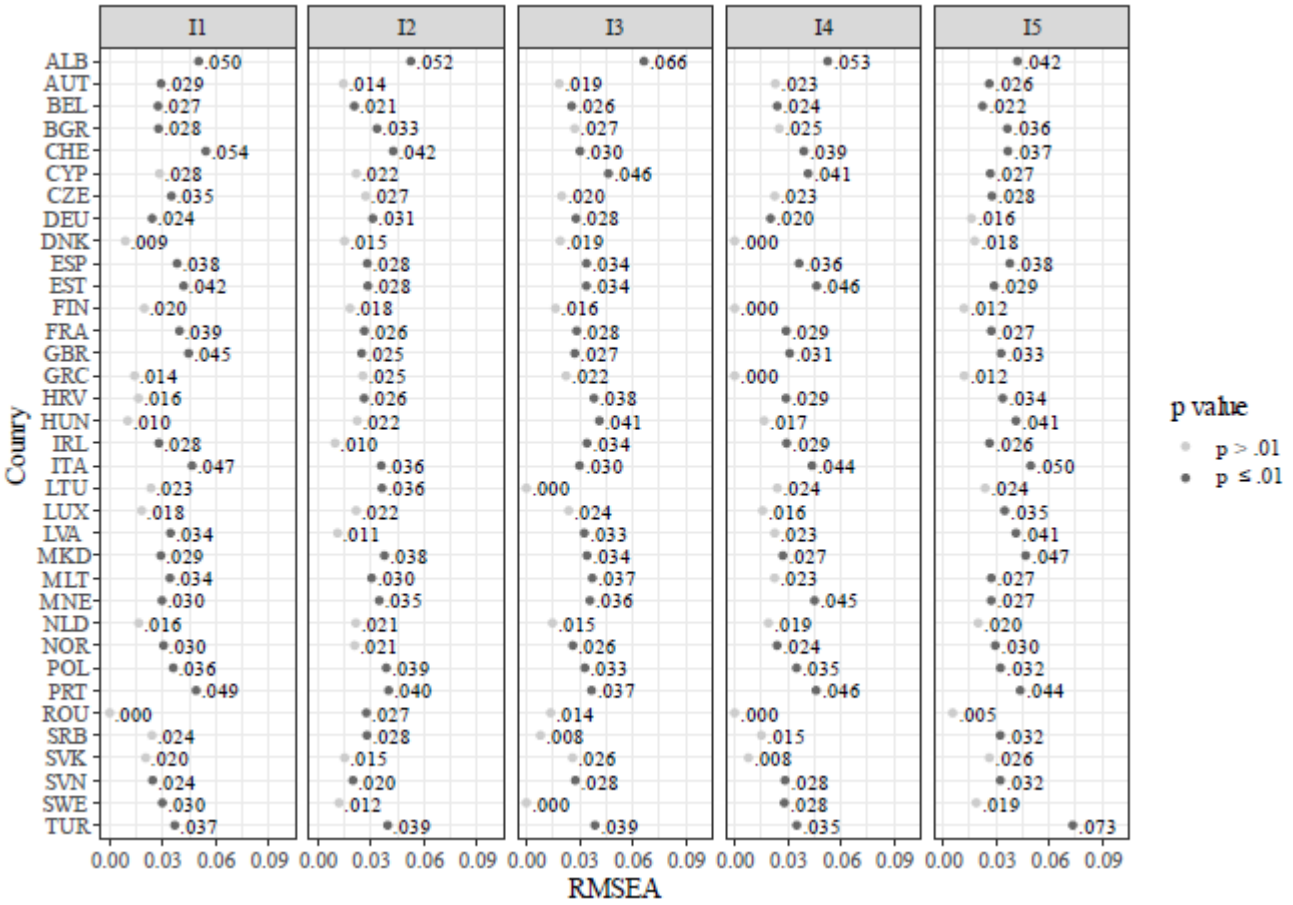


Notes. Values represent the pairwise “folded”/summed Jackknife slope index (JSI) values The cutoff value was determined as the mean of the JSI values plus twice the standard deviation. The red values represent values above the cutoff and are indicative for local dependency.

Appendix

Item fit statistics.

Some items showed some deviation from the GRM. However, the effect sizes are small.

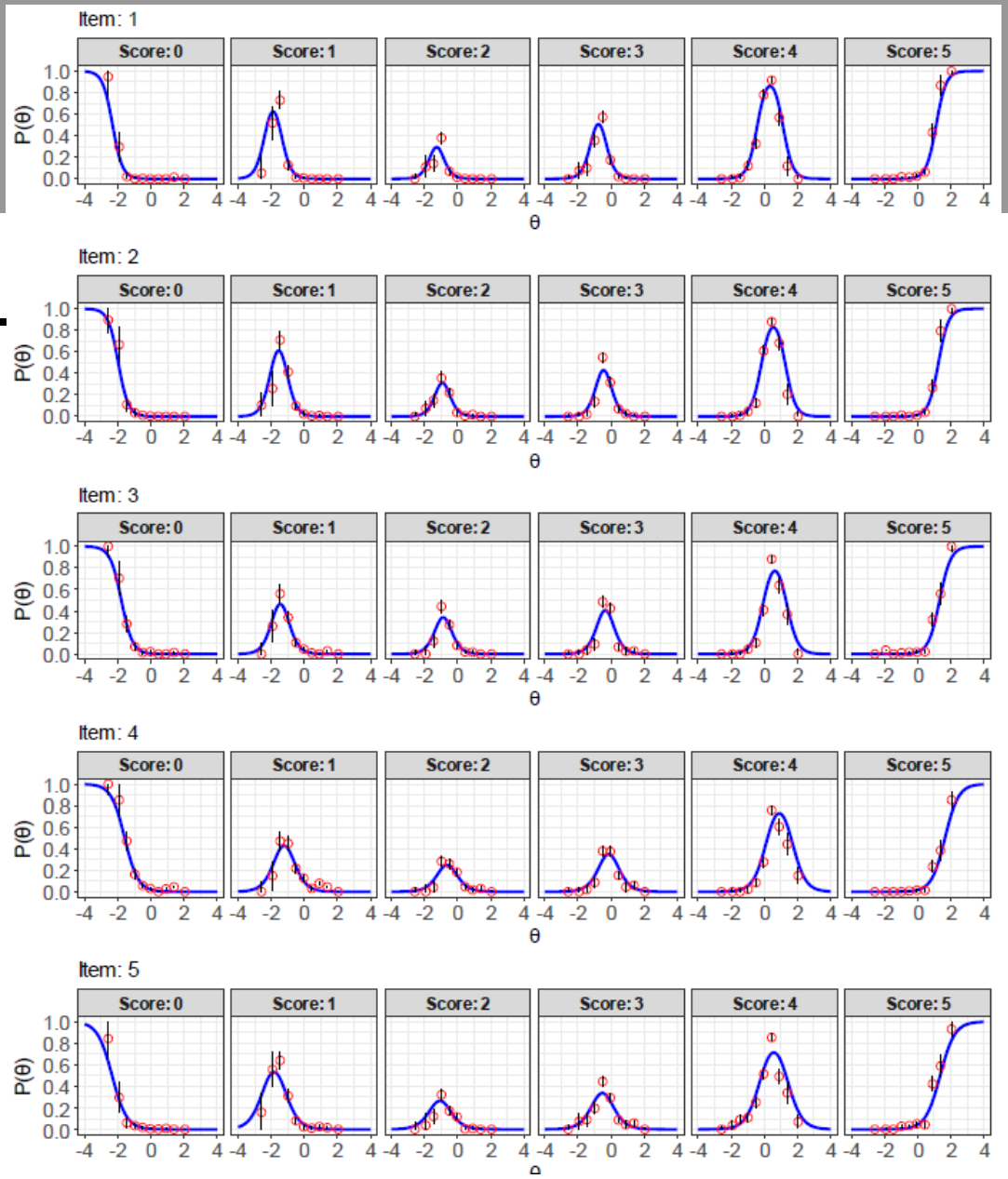


Notes. Values represent RMSEA of item. The significance test is based on the generalized S-X² statistic for polytomous items.

Appendix

Residual plots for GBR.

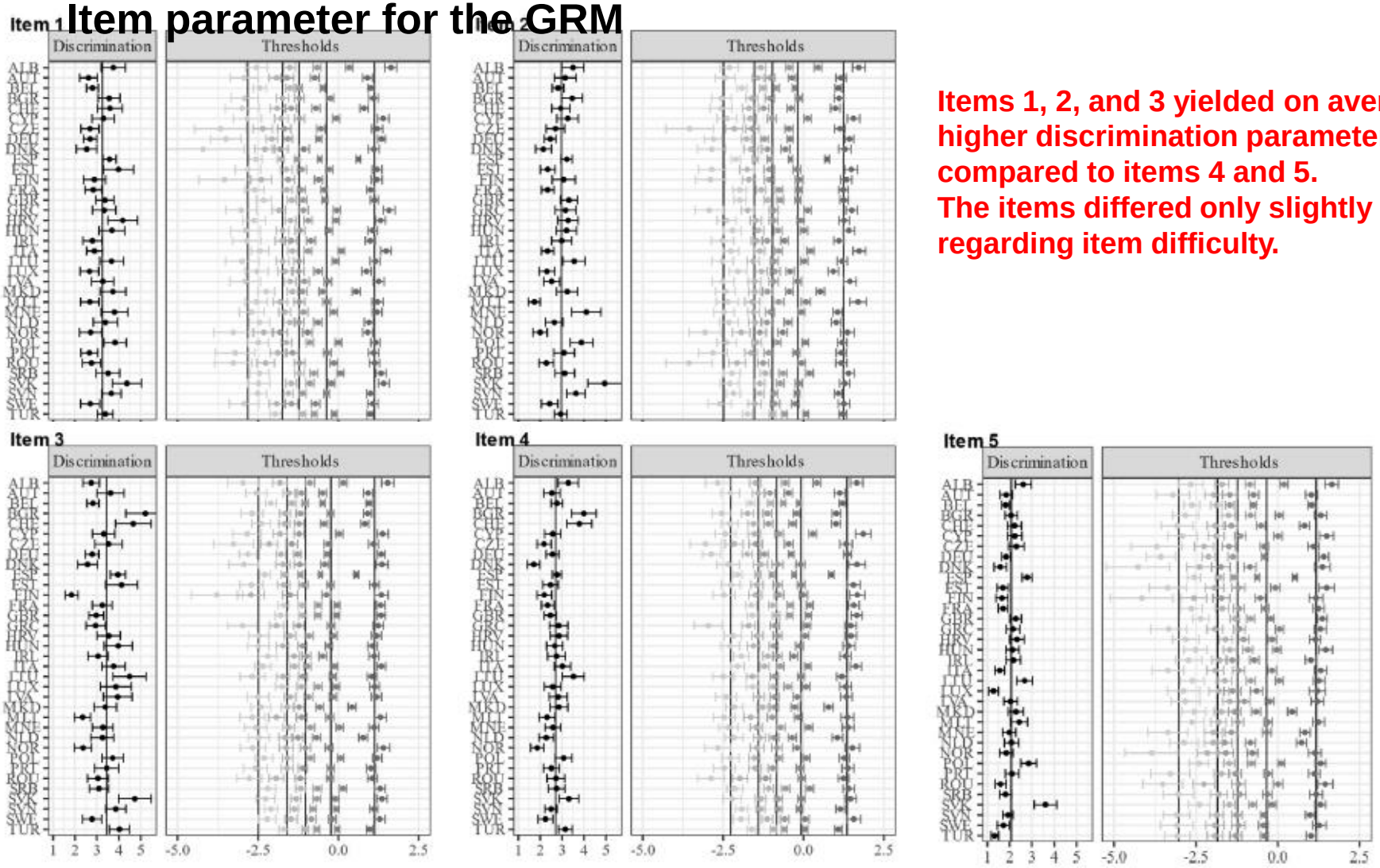
There are only small deviations between observed proportions and response functions.



Notes. Response function and observed proportions with 95% CI.

Appendix

Item parameter for the GRM



Items 1, 2, and 3 yielded on average higher discrimination parameters, compared to items 4 and 5.
The items differed only slightly regarding item difficulty.

Appendix

Reliability.

Test 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.

Table A6. Reliability.

Country	ρ (RMSE)	$\rho_{12.5}$ (SE)	ρ_7 (SE)
ALB	0.92 (0.285)	0.93 (0.272)	0.92 (0.280)
AUT	0.88 (0.338)	0.92 (0.290)	0.90 (0.316)
BEL	0.88 (0.347)	0.91 (0.307)	0.90 (0.312)
BGR	0.93 (0.280)	0.94 (0.237)	0.94 (0.235)
CHE	0.91 (0.300)	0.94 (0.238)	0.94 (0.247)
CYP	0.91 (0.318)	0.91 (0.294)	0.92 (0.287)
CZE	0.89 (0.347)	0.90 (0.315)	0.90 (0.324)
DEU	0.87 (0.365)	0.89 (0.332)	0.88 (0.343)
DNK	0.83 (0.412)	0.86 (0.379)	0.84 (0.398)
ESP	0.90 (0.323)	0.94 (0.252)	0.94 (0.252)
EST	0.89 (0.318)	0.92 (0.278)	0.91 (0.297)
FIN	0.86 (0.372)	0.88 (0.350)	0.86 (0.368)
FRA	0.88 (0.345)	0.90 (0.319)	0.90 (0.321)
GBR	0.90 (0.318)	0.92 (0.280)	0.92 (0.287)
GRC	0.90 (0.312)	0.91 (0.294)	0.91 (0.304)
HRV	0.92 (0.293)	0.93 (0.255)	0.93 (0.270)
HUN	0.91 (0.305)	0.93 (0.271)	0.93 (0.264)
IRL	0.88 (0.350)	0.91 (0.292)	0.90 (0.308)
ITA	0.90 (0.323)	0.91 (0.300)	0.90 (0.309)
LTU	0.92 (0.282)	0.94 (0.241)	0.94 (0.243)
LUX	0.89 (0.342)	0.91 (0.305)	0.90 (0.312)
LVA	0.91 (0.317)	0.92 (0.277)	0.91 (0.297)
MKD	0.90 (0.333)	0.93 (0.260)	0.93 (0.273)
MLT	0.86 (0.369)	0.87 (0.354)	0.88 (0.352)
MNE	0.92 (0.298)	0.93 (0.263)	0.92 (0.275)
NLD	0.88 (0.348)	0.91 (0.292)	0.91 (0.305)
NOR	0.84 (0.403)	0.86 (0.374)	0.85 (0.381)
POL	0.92 (0.280)	0.94 (0.245)	0.94 (0.245)
PRT	0.90 (0.326)	0.91 (0.298)	0.90 (0.315)
ROU	0.88 (0.348)	0.89 (0.333)	0.89 (0.337)
SRB	0.91 (0.308)	0.92 (0.283)	0.90 (0.311)
SVK	0.93 (0.257)	0.96 (0.202)	0.95 (0.212)
SVN	0.91 (0.307)	0.93 (0.260)	0.93 (0.266)
SWE	0.87 (0.364)	0.89 (0.339)	0.88 (0.345)
TUR	0.90 (0.304)	0.93 (0.265)	0.92 (0.275)

Notes. ρ represents the empirical marginal reliability, $\rho_{12.5}$ and ρ7 represents the reliability at the expected test scores 12.5 and 7. RMSE represents the mean root mean square standard error and SE represents the standard error at the respective expected test scores.