

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

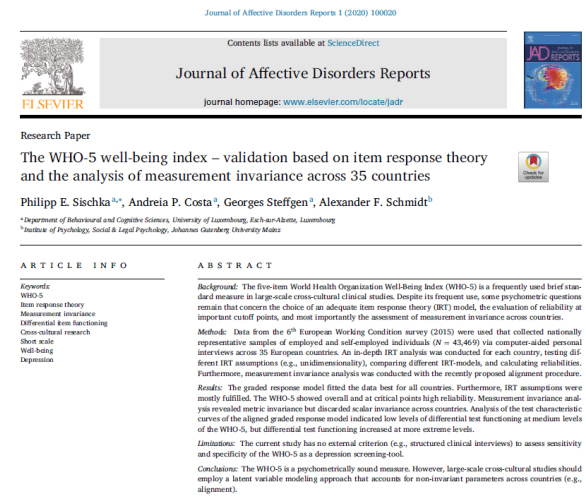


The WHO-5 Well-Being Index – Validation based on item response theory and the analysis of measurement invariance across 35 countries



Session: Various approaches to reducing measurement error

Trier, 2021, July 23th



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 ($M_{\text{sensitivity}} = .86$, $M_{\text{specificity}} = .81$)
 - 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)

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)

Statistical analyses

- Comparing different IRT models (PCM, GPCM, GRM):
 - AIC, BIC (sample-size adjusted)
 - 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)
- IRT analysis
 - Parallel analysis and minimum average partial method (with polychoric correlations; Garrido et al., 2011, 2013)
 - Item/test characteristic curves and information functions, raw residual plots
 - Jackknife Slope Index (JSI; Edwards et al., 2018)
 - Generalized S-X2 item fit index (Kang and Chen, 2011) and RMSEA

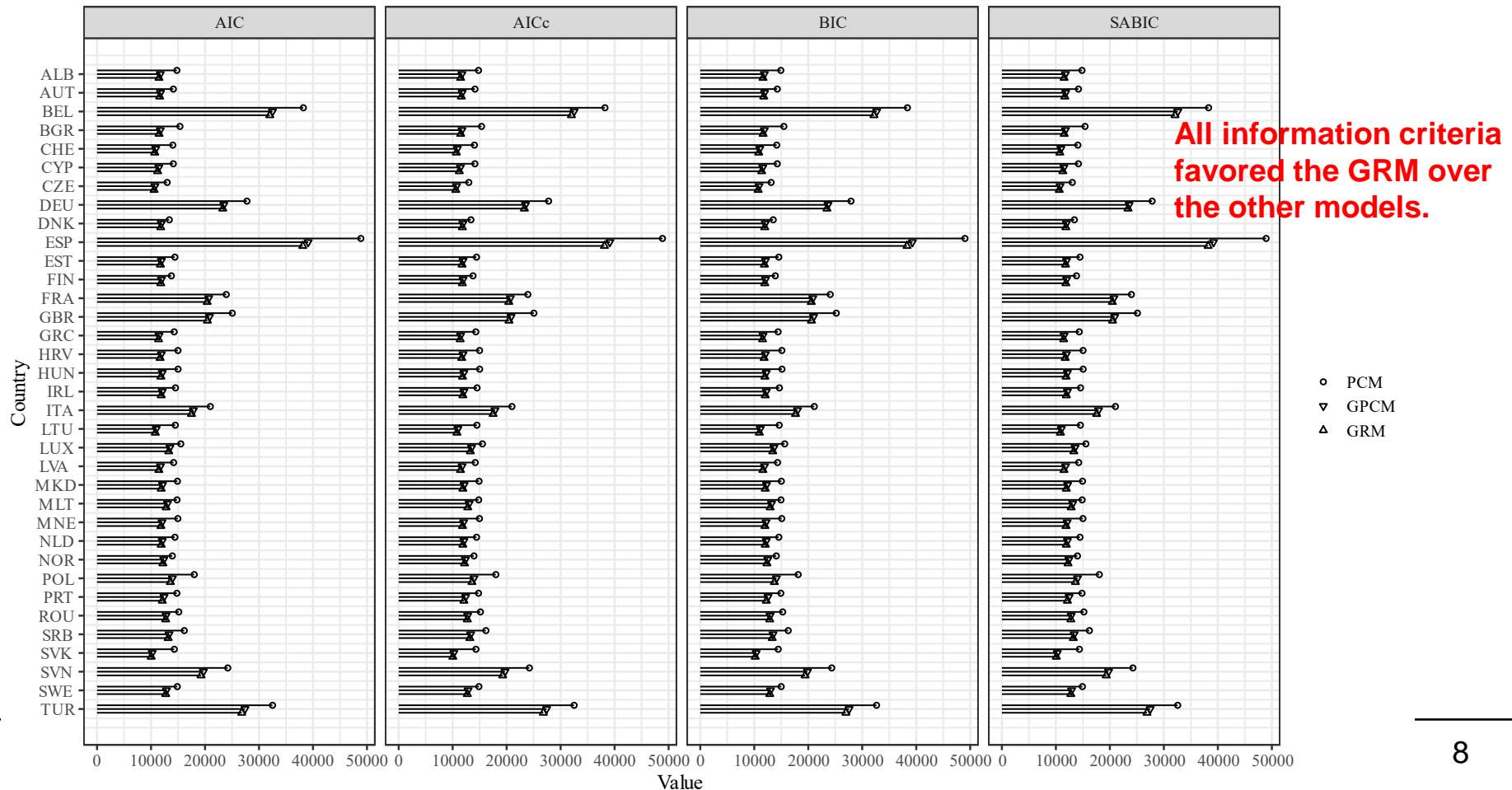
Statistical analyses

- 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.



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

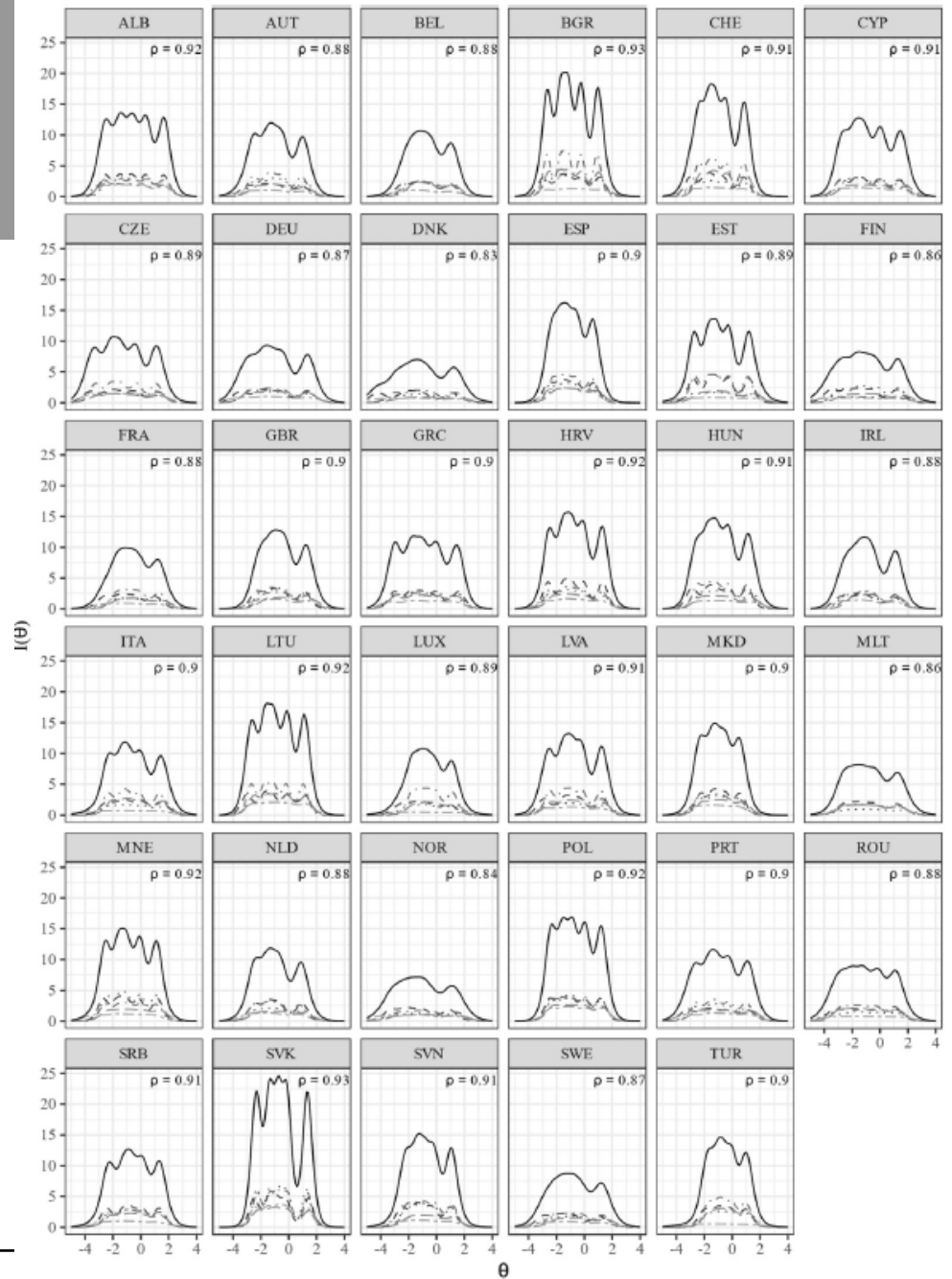
With the exception of some RMSEA values, all fit indices showed adequate fit of the GRM model.

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

Results

Item and test information functions for the GRM.

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



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).

Results

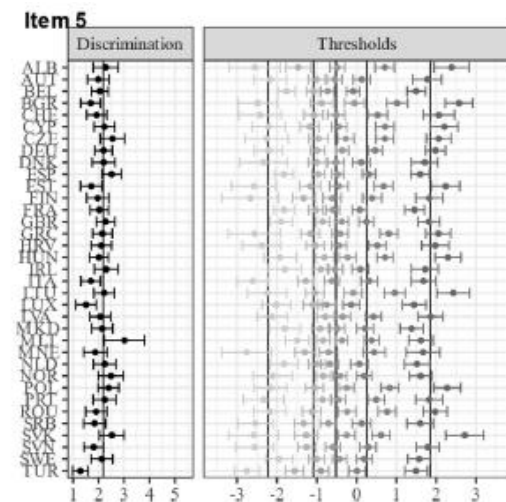
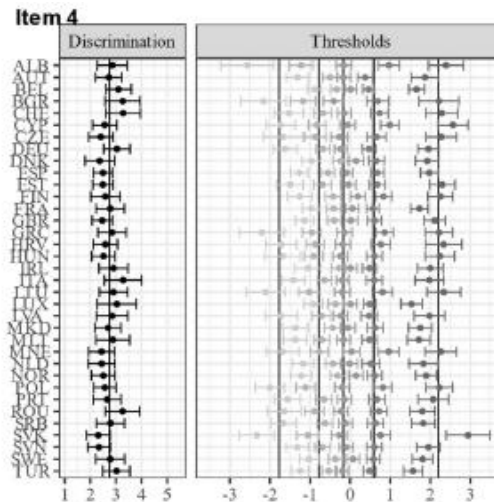
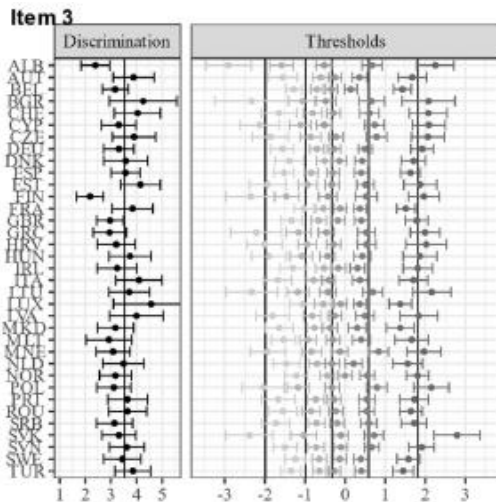
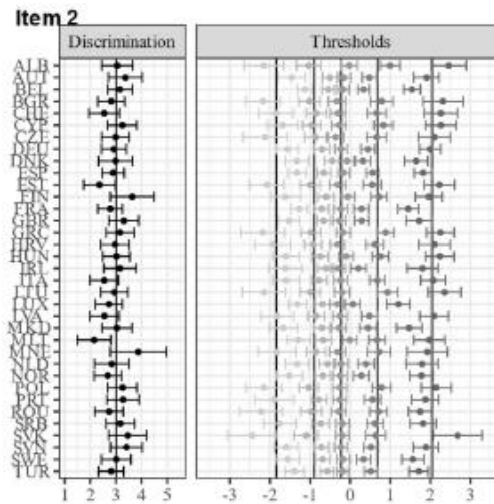
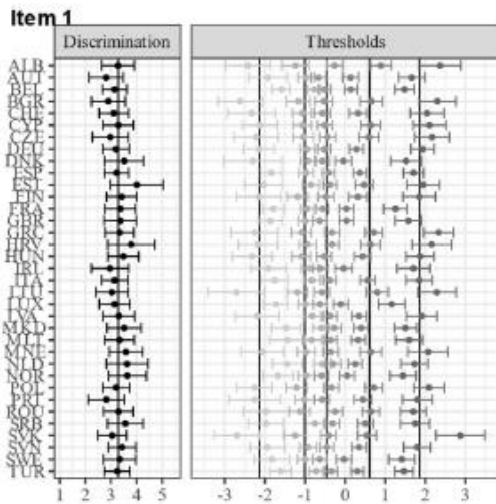
Alignment fit statistics.

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%)
Item 2	Threshold 5	.000	1.86	0.24	1.80	0.33	27 (77.1%)
	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%)
Item 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%)
	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%)
Item 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%)
	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%)
Item 5	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%)
	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%)

Item discrimination parameters nearly invariant (metric invariance). Item threshold parameters for the middle categories showed same amount of invariance, whereas Item threshold parameters for the lower and upper categories are non-invariant.

Results

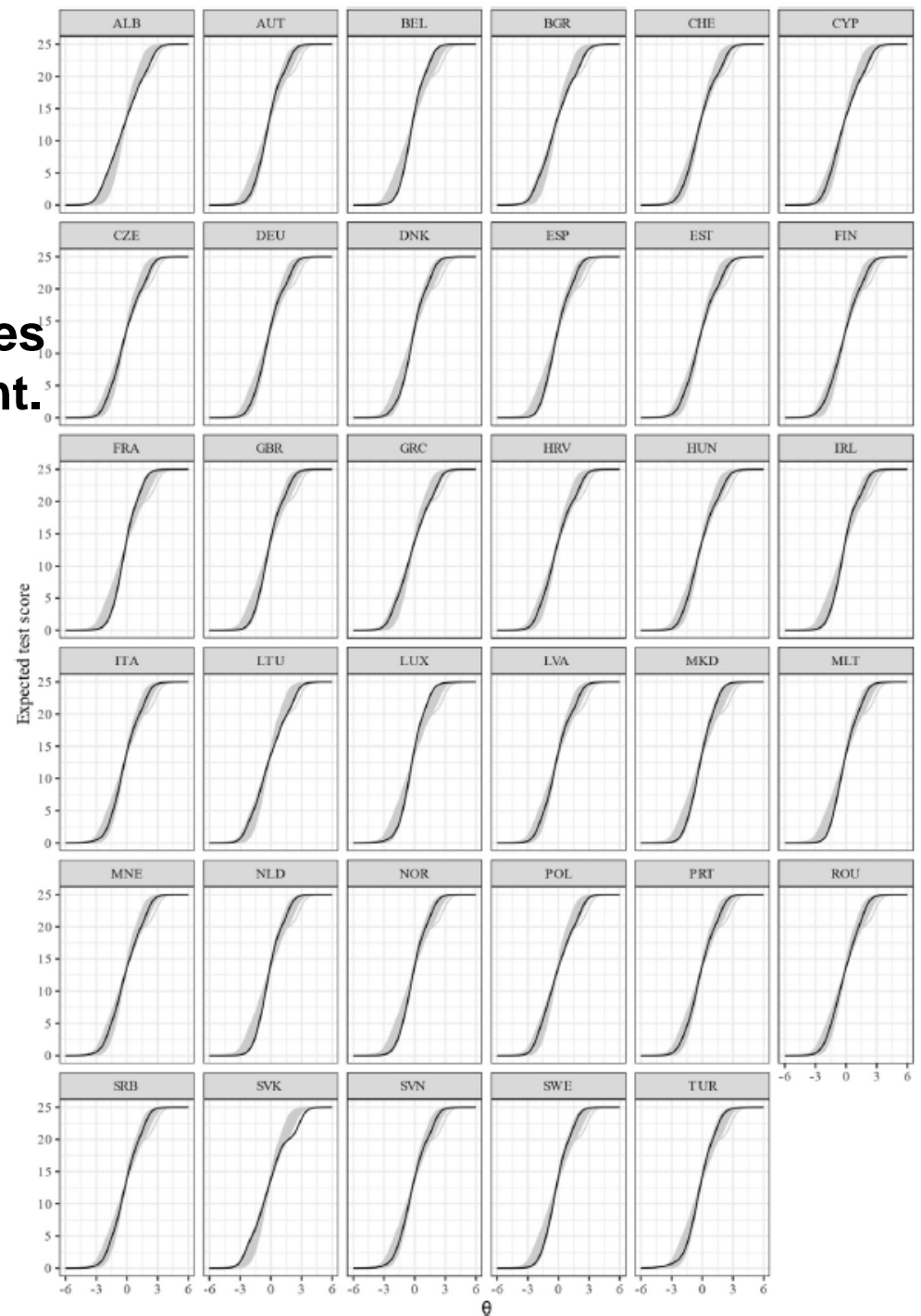
Item parameter for the GRM after alignment.



Item discrimination parameters nearly invariant.
Item threshold parameters for the lower and upper categories are non-invariant.

Results

**Test characteristic curves
for the GRM after alignment.**

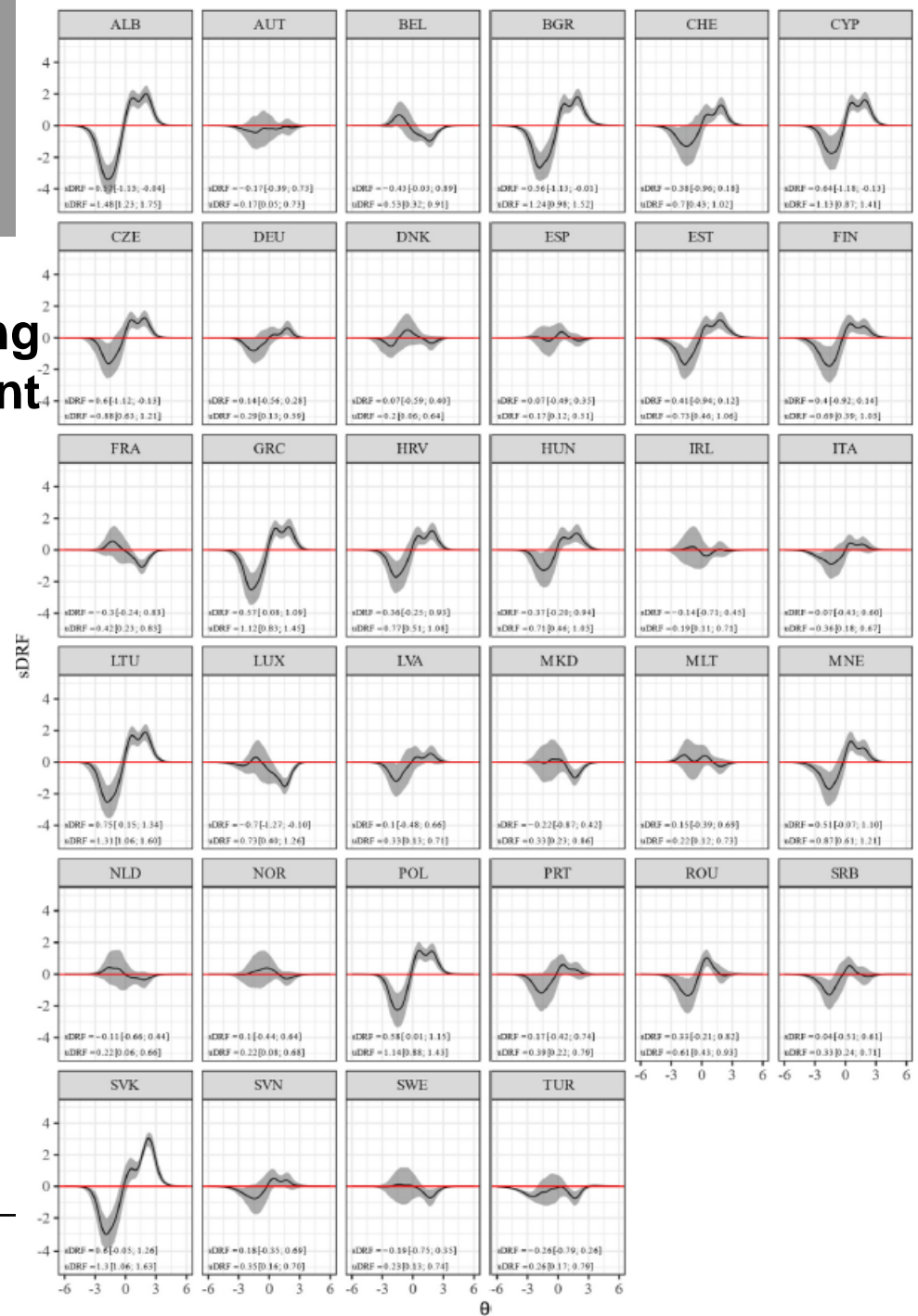


Results

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

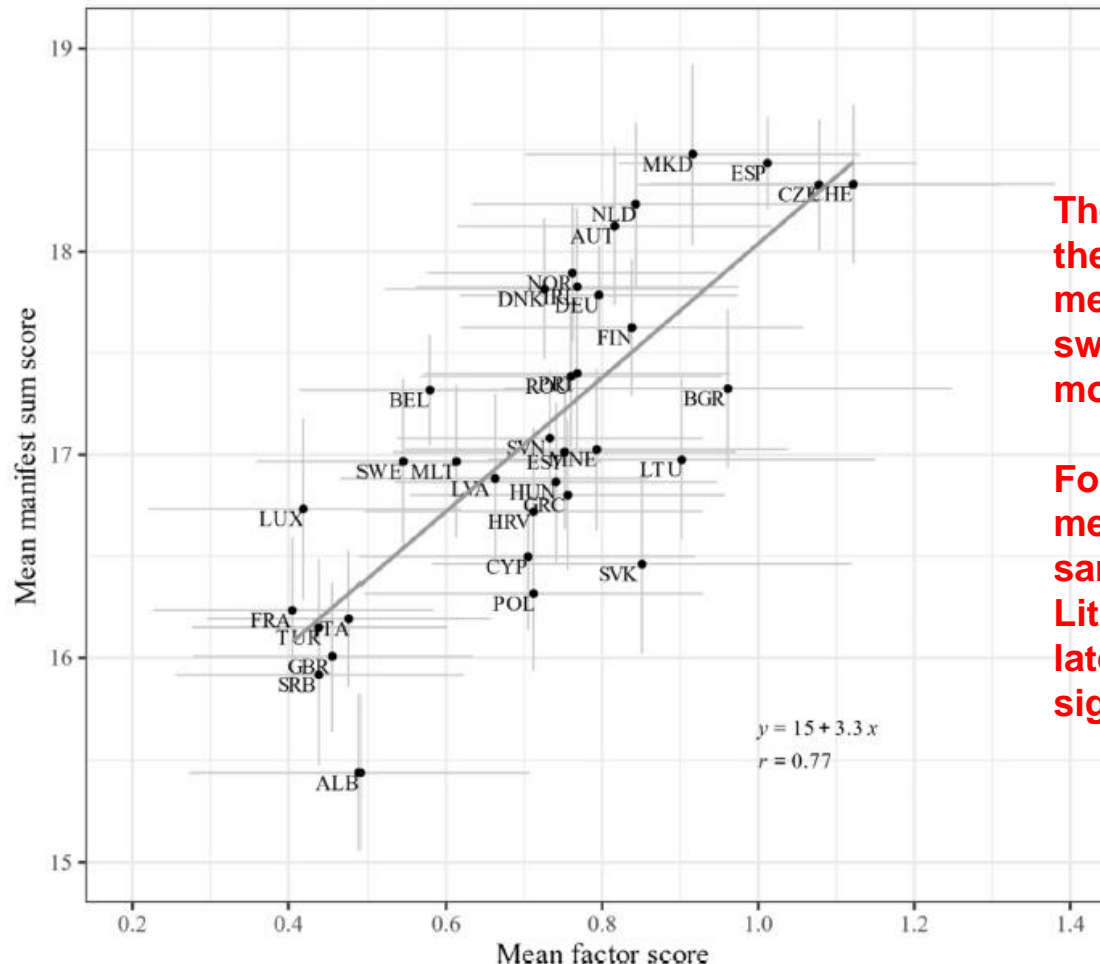
Reading example:

A respondent from Great Britain with a WHO-5 sum score of 2.5 (or 23.1) has the same (estimated) level on the latent well-being variable as a respondent from Albania with a WHO-5 sum score of 5.9 (or 21.1).



Results

Scatterplot with means of factor scores and manifest sum scores.



There is a non-negligible change in the country rank order regarding their mean well-being levels, when switching from manifest to latent models.

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.

Discussion

Summary

- Every criterion favored the GRM over the other IRT models.
- IRT assumptions (unidimensionality, local independence, item fit) could be confirmed.
- Test and item 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 syndrome: 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 index 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 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>.
- 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>.

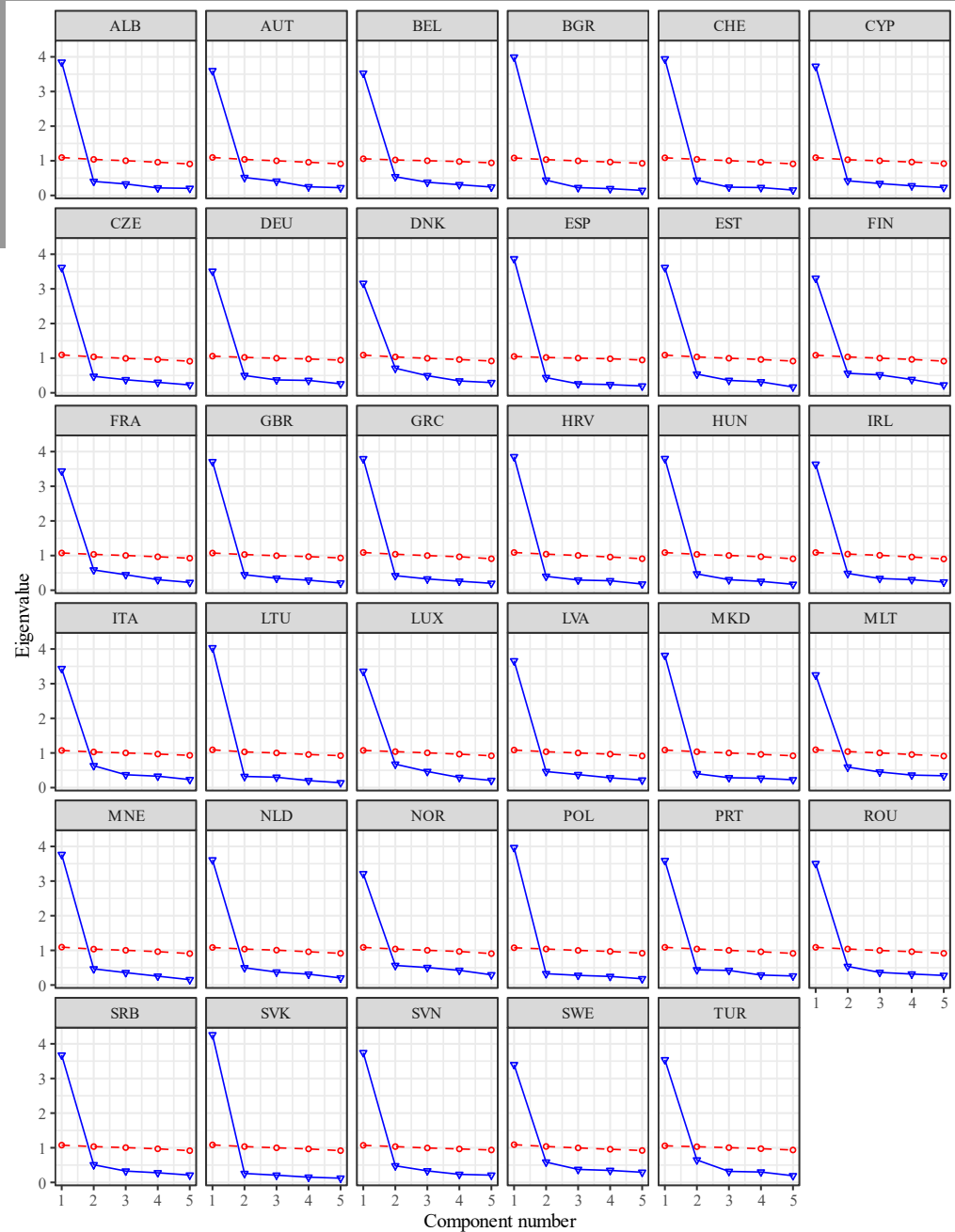


Appendix

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.

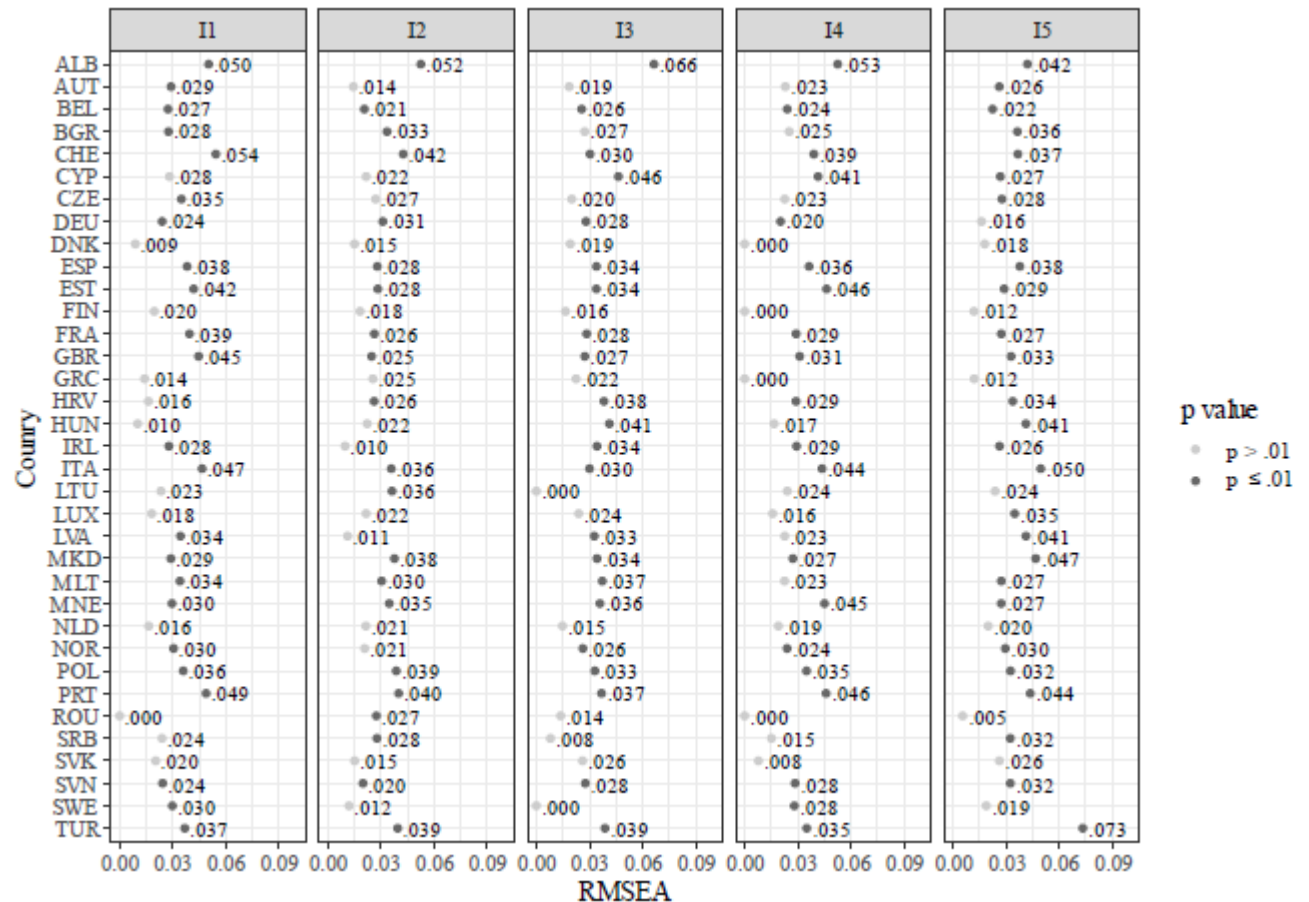
	ALB	AUT	BEL	BGR	CHE	CYP
5	0.17 -3.05 -0.9 -5.2	1.13 0.01 1.29 -0.5	4.09 -1.12 3.97 -2.51	-1.03 -3.79 -2.36 -3.94	0.53 -2.66 -1.21 -1.9	-2.52 -1.91 1.8 -0.15
4	-2.06 0.72 -2.24	-3.75 2.9 2.34	-4.14 8.1 3.38	-4.99 -1.2 0.96	-1.9 1.97 4.72	-1.65 1.45 1.18
3	-2.1 -4.09	-0.92 -1.94	0.26 -6.58	-1.56 -4.44	-1.71 -1.16	-1.55 -2.27
2	2.15 JSI-Cutoff: 2.82	2.51 JSI-Cutoff: 4.58	5.17 JSI-Cutoff: 10.32	-3.33 JSI-Cutoff: 1.15	0.93 JSI-Cutoff: 4.32	4.27 JSI-Cutoff: 4.34
	CZE	DEU	DNK	ESP	EST	FIN
5	-0.02 1.77 0.54 0.69	-0.3 -2.54 -0.42 -3.19	0.99 -3.15 2.11 -3.03	7.31 -2.98 3.76 -3.71	-0.13 -0.26 2.79 0.14	1.07 -0.49 2.88 -0.25
4	-2.5 3.67 -0.66	-5.85 1.55 1.74	-3.41 2.96 2.29	-6.34 7.15 3.97	1.41 5.44 4.2	-3.2 6.12 0.18
3	2.68 -2.76	-1.05 -4.82	-2.87 -3.7	-0.06 -2.23	4.94 -1.83	3.66 -3.64
2	0.56 JSI-Cutoff: 4.48	-0.07 JSI-Cutoff: 3.61	1.99 JSI-Cutoff: 5.1	-1.04 JSI-Cutoff: 10.01	5.66 JSI-Cutoff: 7.69	0.77 JSI-Cutoff: 6.66
	FRA	GBR	GRC	HRV	HUN	IRL
5	4.18 2.19 -1.61 -3.57	1.73 -2.14 0.76 -1.48	3.11 -0.29 0.7 -0.12	1.31 1.39 -2.06 0.1	2.77 -1.54 1.37 -2.93	0.24 -2.14 0.59 -1.63
4	-7 -2.27 13.02	-4.48 1.11 5.35	-1.96 1.92 4.14	-3.88 0.54 1.21	-2.81 2.3 -0.23	-3.26 1.83 3.54
3	-5.65 -5.78	-2.59 -2.53	-1.15 -1.83	1.49 -3.92	2.44 -4.59	-1.77 -1.97
2	8.59 JSI-Cutoff: 13.54	5.37 JSI-Cutoff: 6.85	4.35 JSI-Cutoff: 5.63	2.27 JSI-Cutoff: 4.43	1.84 JSI-Cutoff: 5.2	-0.58 JSI-Cutoff: 3.64
	ITA	LTU	LUX	LVA	MKD	MLT
5	1.19 -3.3 1.05 -2.65	-0.22 -1.24 -3.86 -1.47	0.6 -1.64 -1.3 -1.83	-0.57 -1.22 -1.44 -3.64	-0.48 -1.31 -0.04 -1.42	0.85 -1.53 0.81 0
4	-5.27 0.78 4.26	-3.83 -2.41 2.58	-5.65 -2.59 8.81	-4.07 -1.65 1.03	-0.22 -1.18 -1.22	-2.71 0.11 4.92
3	0.19 -3.49	-2.29 -3.87	-4.12 -3.86	-2.04 -2.01	-2.21 -1.55	-2.64 -3.11
2	-0.96 JSI-Cutoff: 4.89	2.8 JSI-Cutoff: 3.55	3.03 JSI-Cutoff: 7.57	0.13 JSI-Cutoff: 1.56	-0.5 JSI-Cutoff: 0.34	0.07 JSI-Cutoff: 4.43
	MNE	NLD	NOR	POL	PRT	ROU
5	-3.16 -3.6 -1.34 -2.22	-0.49 -2.06 1.78 -0.91	2.59 0.26 -0.88 -0.96	0.83 -0.91 0.6 1.25	-1.43 0.45 -0.88 -0.61	0.01 -1.23 1.18 0.63
4	-5.85 -1.07 2.02	-4.02 0.78 0.65	-4.11 2.54 4.3	-2.24 3.25 0.6	-4.05 -0.68 2.48	0.29 1.64 3.22
3	-3.73 -5.4	-0.97 -4.78	-0.49 -1.1	-0.21 -4.47	-0.81 -2.51	-0.34 0.48
2	5.55 JSI-Cutoff: 5.03	2.7 JSI-Cutoff: 4.05	2.28 JSI-Cutoff: 5.39	1.8 JSI-Cutoff: 4.4	1.41 JSI-Cutoff: 3.05	0.42 JSI-Cutoff: 3.06
	SRB	SVK	SVN	SWE	TUR	
5	-2.9 -2.5 -0.45 -1.69	1.53 0.71 3.63 3.99	-1.91 -0.71 -1.8 -2.32	1.35 -2.28 2.23 -2.04	1.54 2.63 1.55 1.21	1 2 3 4
4	-3.77 0.87 2.65	-0.83 0.42 2.71	-3.95 0.12 0.12	-3.28 3.21 1.19	-4.92 4.55 4.8	
3	-2.37 -2.95	0.22 0.96	0.29 -2.27	0.59 -2.82	4.65 -3.34	
2	2.32 JSI-Cutoff: 3.53	6 JSI-Cutoff: 6.13	0.15 JSI-Cutoff: 1.64	-0.2 JSI-Cutoff: 4.34	2.1 JSI-Cutoff: 8.02	

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.

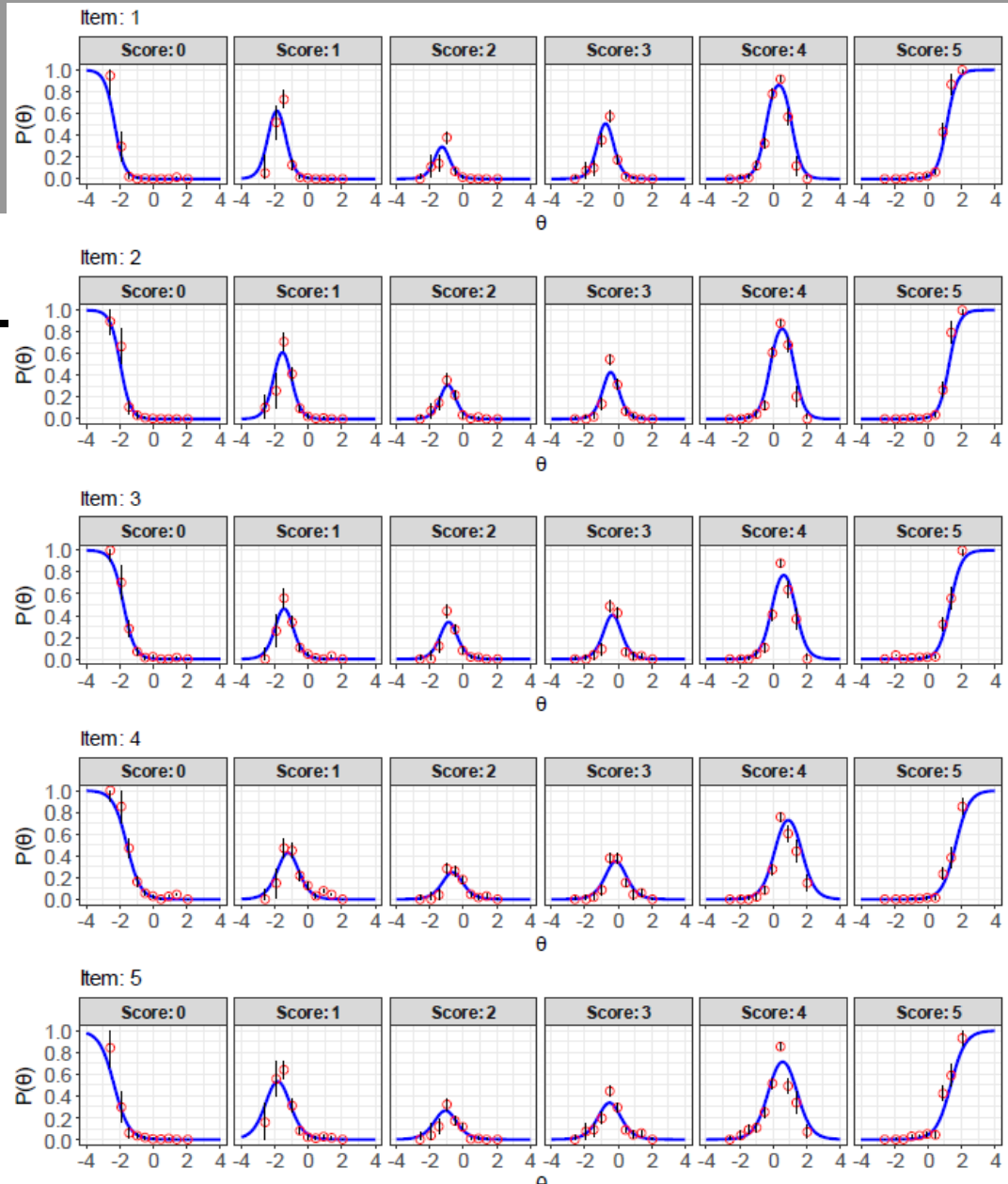


Notes. Values represent RMSEA of item. The significance test is based on the generalized S-X² statistic for polytomous items.

Appendix

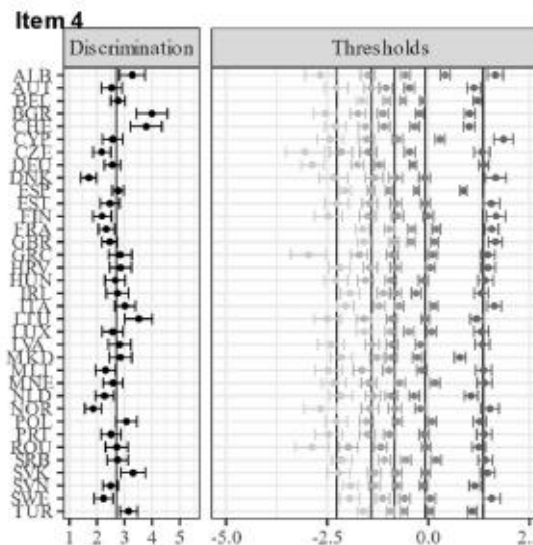
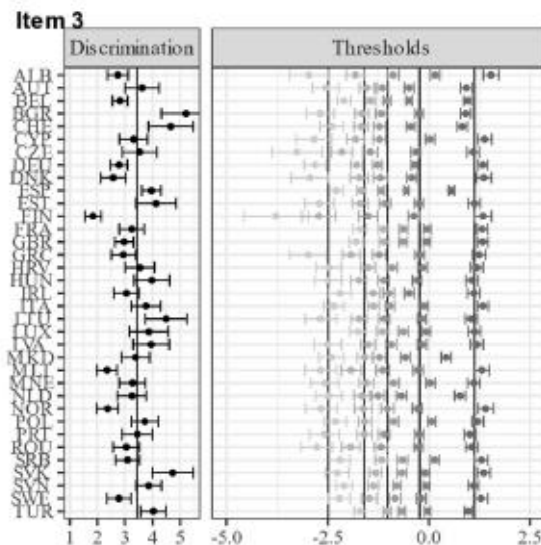
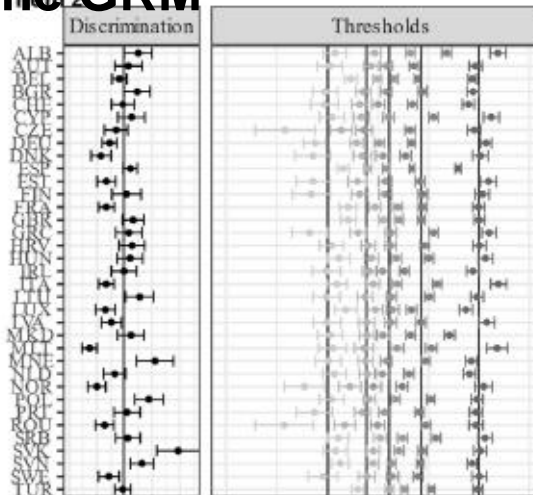
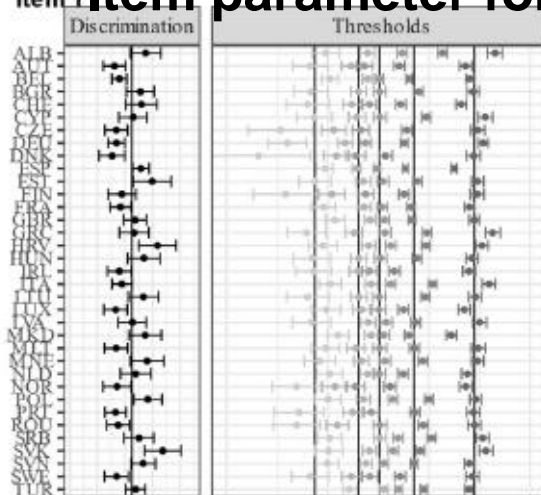
Residual plots for GBR.

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

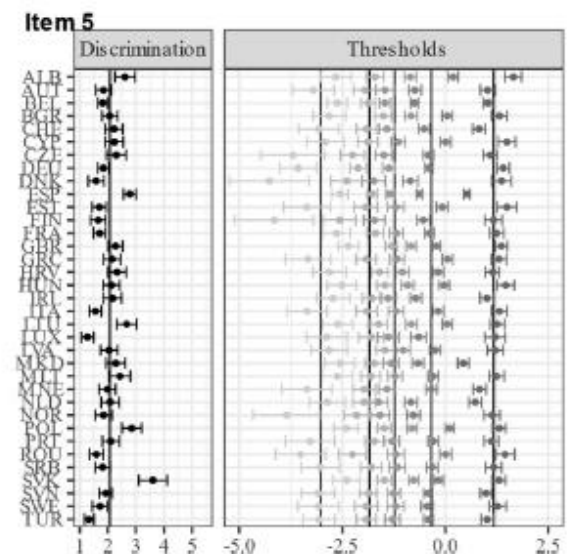


Appendix

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 ($\rho = .83-.93$) as well as at critical points ($\rho_{12.5} = .86-.96$, $\rho_7 = .84-.95$) 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 ρ_7 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.