Full Psychometric Audit: Technical Appendix

Measurement Quality Audit, Tier 2. Sample appendix for an eight-item Programme Climate Scale administered in three language versions, prepared by Dr Milos Kankaras, Center for Psychology, September 2026.

This is a demonstration on synthetic data. The dataset was simulated with one known defect built in (the wording of item 5 in Language 3 reads milder than the source) so that the reader can see the analysis find it. Every number below is the actual output of the analysis on that data; nothing is illustrative. A client receives this appendix on their own data, with the same tables.

1. Verdict

The scale is unidimensional and reliable in all three languages, and its loadings are equivalent across them, so the construct is the same thing in each. Scores are not directly comparable across languages as they stand: full scalar invariance fails, and the failure is located in one item. Item 5 in Language 3 is answered more positively than its loading warrants, by about half a scale point. With that item's thresholds freed, partial scalar invariance holds and latent means can be compared. On that comparison Language 3 scores lower than the source by 0.19 standard deviations, a difference the raw totals understate because the item 5 shift pulls Language 3 up. The fix is a re-translation of item 5 before the next wave, and in the meantime reporting group comparisons from the partial-invariance model rather than from totals.

2. Data

One file, 1,230 respondents (source language 450, Language 2 400, Language 3 380), eight items on a five-point agreement scale, no missing data, one administration. Items were treated as ordered categorical; a continuous-estimator run was kept as a sensitivity check (section 6).

3. Item analysis

Table 1 gives item means, standard deviations, the share at the ceiling category and corrected item-total correlations per language. Item 5 in Language 3 stands out: its mean is the highest in that group while the other seven items are the lowest of the three languages, and its ceiling share is twice that of the source.

Table 1. Item descriptives per language.

Item	EN M	EN SD	EN top %	EN r(it)	L2 M	L2 SD	L2 top %	L2 r(it)	L3 M	L3 SD	L3 top %	L3 r(it)
pcs1	3.42	0.97	14	0.66	3.42	0.97	12	0.67	3.24	0.98	9	0.66
pcs2	3.22	1.01	10	0.66	3.26	1.02	12	0.63	3.12	1.02	9	0.66
pcs3	3.52	0.96	16	0.68	3.58	0.96	20	0.65	3.38	1.03	15	0.68

Item	EN M	EN SD	EN top %	EN r(it)	L2 M	L2 SD	L2 top %	L2 r(it)	L3 M	L3 SD	L3 top %	L3 r(it)
pcs4	3.03	1.05	8	0.56	3.18	1.03	11	0.61	2.92	1.02	6	0.55
pcs5	3.09	0.99	7	0.61	3.25	0.99	9	0.59	3.57	0.96	17	0.63
pcs6	3.32	0.97	11	0.67	3.37	1.01	14	0.60	3.21	1.00	10	0.68
pcs7	2.94	0.96	4	0.56	2.96	0.95	5	0.48	2.78	0.99	4	0.52
pcs8	3.26	1.00	8	0.63	3.29	1.02	12	0.63	3.10	0.98	7	0.55

4. Reliability and single-group fit

Internal consistency is good and near-identical across languages. The one-factor model fits well within each language on its own (Table 2), which is the configural condition the invariance cascade starts from.

Table 2. Reliability and one-factor CFA fit per language (WLSMV, scaled indices).

Group	n	alpha	omega	chi2	df	CFI	TLI	RMSEA [90% CI]	SRMR
Source (EN)	450	0.872	0.873	17.71	20	1.000	1.001	0.000 [0.000, 0.036]	0.019
Language 2	400	0.861	0.862	21.83	20	0.999	0.999	0.015 [0.000, 0.047]	0.024
Language 3	380	0.866	0.867	19.78	20	1.000	1.000	0.000 [0.000, 0.044]	0.022

Table 3. Standardised loadings per language (single-group models).

Item	Source (EN)	Language 2	Language 3
pcs1	0.75	0.77	0.75
pcs2	0.74	0.72	0.74
pcs3	0.78	0.76	0.78
pcs4	0.63	0.69	0.63
pcs5	0.69	0.67	0.72
pcs6	0.76	0.69	0.76
pcs7	0.64	0.55	0.59
pcs8	0.71	0.71	0.63

5. Measurement invariance across languages

Table 4 reports the cascade with ordinal (WLSMV) estimation, thresholds and loadings constrained together at the metric step as the ordinal cascade requires. Decision rule: a more constrained model is retained when the change in CFI is no larger than .010 and the change in RMSEA no larger than .015 (Chen, 2007; Cheung & Rensvold, 2002). Metric invariance holds. Scalar invariance is rejected, and the score test locates the misfit in the thresholds of item 5 in Language 3. Freeing those thresholds restores fit to the metric level; that is the partial scalar model retained for group

comparison. Strict invariance is rejected, which affects nothing here: it is not required for comparing means.

Table 4. Tests of measurement invariance across language (WLSMV, N = 1,230).

Model	chi2	df	CFI	RMSEA [90% CI]	dchi2 (ddf)	dCFI	dRMSEA	Decision
M1 Configural	59.46	60	0.999	0.012 [0.000, 0.038]	-	-	-	reference
M2 Metric (loadings)	109.59	106	1.000	0.009 [0.000, 0.027]	49.16 (46)	+0.000	-0.003	supported
M3 Scalar (loadings, intercepts)	285.63	120	0.978	0.058 [0.049, 0.067]	139.19 (14)	-0.021	+0.049	rejected
M4 Partial scalar (item 5 thresholds freed in L3)	120.21	114	0.999	0.012 [0.000, 0.028]	9.21 (8)	-0.000	+0.002	partial
M5 Strict (plus residuals)	292.11	136	0.979	0.053 [0.045, 0.061]	151.73 (22)	-0.020	+0.041	rejected

Note. Scaled chi-square and difference tests. dCFI and dRMSEA are relative to the last retained model: M2 vs M1, M3 vs M2, M4 vs M2, M5 vs M4. SRMR: .022, .025, .025, .025, .030 for M1 to M5. Decision: reference, supported, rejected, partial.

6. Sensitivity: continuous estimation

Treating the five-point items as continuous (MLR) gives the same verdict: metric holds, scalar is rejected (dCFI -0.036), and freeing the intercept of item 5 restores partial scalar fit (dCFI +0.000 against metric). The ordinal run is the one reported because the items have five categories; the agreement between the two is why the verdict is stated without qualification.

7. Where invariance fails, and by how much

Table 5 gives the item 5 thresholds per language from the configural model. In Language 3 every threshold sits lower than in the source, by 0.6 to 0.8 units on the latent response scale: a respondent at the same level of the construct crosses each category boundary more easily, which is what a milder translation produces. The pattern is uniform across thresholds, so it is a shift rather than a change in discrimination, and the loading (Table 3) is unaffected.

Table 5. Item 5 thresholds per language, configural model.

Group	t1	t2	t3	t4
Source (EN)	-2.30	-0.83	0.56	2.02
Language 2	-2.35	-0.98	0.23	1.82
Language 3	-3.01	-1.60	-0.21	1.40

8. Group comparison

Under the partial scalar model, with the source language as reference, the latent mean of Language 2 is +0.08 (SE 0.07, p = 0.29) and of Language 3 is -0.19 (SE 0.07, p = 0.009). Language 3 scores lower on the construct; Language 2 does not differ. The observed totals (source 25.8, Language 2 26.3,

Language 3 25.3) show the same ordering with a smaller Language 3 gap, because item 5 contributes about half a point to Language 3's total that the construct does not explain. Any report built on totals would have understated the Language 3 difference.

9. Fix-list

Band	Action	Type
High	Re-translate item 5 in Language 3 against the source, with a back-translation, before the next wave; the current wording reads milder than the source by about half a scale point.	Wording
High	Until then, report cross-language comparisons from the partial scalar model, not from totals, and say that item 5 is freed in Language 3.	Analysis
Medium	Re-run the cascade after the re-translation on the next wave's data to confirm full scalar invariance; the scripts supplied make this a one-line change.	Analysis
Low	Item 7 has the lowest loading in every language (.55 to .64) and the lowest ceiling; it is not a defect, but it is the first item to review if the scale is ever shortened.	Wording

10. Methods

Confirmatory factor models were estimated in R 4.5.3 with lavaan 0.6.21 and semTools 0.5.8. Ordinal models used the WLSMV estimator with theta parameterisation and the Wu and Estabrook (2016) identification, with thresholds and loadings constrained together at the metric step; the sensitivity run used MLR with full-information maximum likelihood. Fit was evaluated with the scaled CFI, TLI, RMSEA with its 90 percent interval and SRMR. Invariance decisions followed Chen (2007): dCFI within .010 and dRMSEA within .015 relative to the last retained model; the scaled chi-square difference is reported but not used as the criterion. When a level failed, the score test identified the worst constraint and it was released one at a time, to a maximum of five, until the fit criterion was met. Reliability is coefficient alpha and McDonald's omega from the single-group continuous model. The model settings, the data file and the analysis code are supplied with this appendix so every number can be re-run.

References

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling, 14(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233-255. https://doi.org/10.1207/s15328007sem0902_5

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. Psychometrika, 81(4), 1014-1045. https://doi.org/10.1007/s11336-016-9506-0

Auditor's verdict

The scale measures the same construct in all three languages and its scores can be compared across them once item 5 in Language 3 is either re-translated or freed in the model. Comparisons made on raw totals as the scale currently stands understate the Language 3 difference and should not be reported.

Dr Milos Kankaras, Founder and Director, Center for Psychology

Specimen: a client report carries the reviewer's dated signature here.

Dr Milos Kankaras, Founder and Director, Center for Psychology | milos@centerforpsychology.me | miloskankaras.com | centerforpsychology.me | ORCID | Google Scholar | LinkedIn