Online Supplement

Table 1A

Item Parameters

Blocks	Items				Pairwise comparisons	
#	#	Attribute	Loading	var (error)	#	threshohold
1	1	N	-0.705	1.463	{1,2}	0.231
	2	E	1.108	0.242	{1,3}	1.287
	3	O	1.024	1	{2,3}	1.161
2	4	A	0.845	1.004	{4,5}	-0.327
	5	C	0.994	0.287	{4,6}	-1.237
	6	N	-0.804	1	{5,6}	-0.834
3	7	O	-0.476	0.052	{7,8}	0.203
	8	E	0.624	2.904	{7,9}	1.999
	9	A	0.822	1	{8,9}	2.102
4	10	C	0.734	1.083	{10,11}	0.415
	11	O	0.974	0.466	{10,12}	-2.155
	12	N	0.722	1	{11,12}	-2.507
5	13	A	0.552	1.249	{13,14}	-2.144
	14	N	1.194	3.157	{13,15}	-0.151
	15	E	1.243	1	{14,15}	1.813
6	16	O	0.903	0.92	{16,17}	-1.803
	17	E	-0.719	1.091	{16,18}	-2.514
	18	C	-0.720	1	{17,18}	-0.71
7	19	E	-1.251	2.711	{19,20}	-1.575
	20	N	1.864	4.828	{19,21}	2.378
	21	A	0.607	1	{20,21}	4.524
8	22	C	0.667	0.56	{22,23}	0.034
	23	O	0.665	0.287	{22,24}	-1.462
	24	E	-0.698	1	{23,24}	-1.505
9	25	O	1.235	4.803	{25,26}	-2.922
	26	N	1.379	2.276	{25,27}	-3.404
	27	A	-1.116	1	{26,27}	-0.499
10	28	C	-0.821	0.629	{28,29}	0.358
	29	N	0.636	0.675	{28,30}	1.636
	30	E	1.141	1	{29,30}	1.408
11	31	E	0.847	0.458	{31,32}	0.29
	32	A	0.676	0.42	{31,33}	0.165
	33	C	0.838	1	{32,33}	-0.249
12	34	N	-0.455	1.219	{34,35}	0.629

	35	A	0.570	0.79	{34,36}	-1.132
	36	O	-0.787	1	{35,36}	-1.887
13	37	E	-0.830	0.731	{37,38}	-0.673
	38	N	1.004	0.756	{37,39}	-0.598
	39	C	-0.985	1	{38,39}	-0.03
14	40	A	0.798	1.356	{40,41}	0.112
	41	C	1.114	1.134	{40,42}	0.301
	42	O	1.158	1	{41,42}	0.486
15	43	E	0.838	0.937	{43,44}	-0.178
	44	O	0.983	0.855	{43,45}	-2.307
	45	N	1.109	1	{44,45}	-2.424
16	46	C	1.202	2.074	{46,47}	-1.192
	47	N	-1.115	3.062	{46,48}	-3.287
	48	A	-0.417	1	{47,48}	-2.363
17	49	C	-0.830	2.558	{49,50}	2.644
	50	A	0.782	1.696	{49,51}	3.038
	51	O	0.878	1	{50,51}	0.775
18	52	A	-0.899	3.309	{52,53}	2.844
	53	E	0.956	3.941	{52,54}	-0.227
	54	O	-0.702	1	{53,54}	-2.758
19	55	O	-0.463	1.823	{55,56}	1.929
	56	C	0.801	1.036	{55,57}	-0.683
	57	N	0.701	1	{56,57}	-2.444
20	58	C	0.546	1.167	{58,59}	-1.992
	59	A	-0.339	0.315	{58,60}	-0.031
	60	E	1.052	1	{59,60}	2.027

Note. N = Neuroticism; E = Extroversion; O = Openness; A = Agreeableness; C = Conscientiousness. Error variance of last item in each block is set to 1 for identification. From Modeling Forced-Choice Response Format (pp. 547 –848). In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test* (pp. 523-569), by A. Brown and A. Maydeu-Olivares, 2018, London, UK: John Wiley & Sons Ltd.

Table 2A

Nonconvergence Rate

	CI_S		CI_M		MI_S		MI_M		SI_S		SI_M	
	5i	10i	5i	10i	5i	10i	5i	10i	5i	10i	5i	10i
Null	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
PHL	.02	.00	.01	.00	.00	.00	.00	.00				
NHL	.00	.00	.00	.00	.00	.00	.00	.00				
PLL	.00	.00	.00	.00	.00	.00	.00	.01				
NLL	.02	.00	.10	.00	.00	.00	.06	.00				
PHT	.01	.00	.01	.02	.00	.00	.00	.00	.00	.00	.00	.00
NHT	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
PLT	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
NLT	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note. See Table 1 for the condition names. CI = Configural Invariance; MI = Metric Invariance; SI = Scalar Invariance; S = small magnitude of non-invariance; M = medium magnitude of non-invariance; Null = Measurement invariance.

Table 3A

Proportion of Poor Model Fit: CFI < .95

	CI_S		CI_M		MI_S		MI_M		SI_S		SI_M	
	5i	10i	5i	10i	5i	10i	5i	10i	5i	10i	5i	10i
Null	.01	.01	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00
PHL	.01	.00	.00	.00								
NHL	.00	.01	.00	.00								
PLL	.01	.01	.00	.02								
NLL	.02	.02	.01	.01								
PHT	.00	.01	.00	.00	.00	.00	.00	.00				
NHT	.00	.00	.01	.03	.00	.00	.01	.00				
PLT	.01	.01	.01	.02	.00	.00	.00	.00				
NLT	.01	.00	.00	.01	.00	.00	.00	.00				

Note. See Table 1 for the condition names.

Table 4A
Absolute Bias, Bias and RMSE

	Abs_Bias_S		**Abs_Bias_M**		**Bias_S**		**Bias_M**		**RMSE_S**		**RMSE_M**	
	5i	10i	5i	10i	5i	10i	5i	10i	5i	10i	5i	10i
PHL	.354	.340	.338	.344	-.010	.002	-.025	-.004	1.246	1.270	1.175	1.260
NHL	.359	.355	.384	.354	.010	.002	.025	-.001	1.304	1.277	1.136	1.258
PLL	.342	.342	.354	.368	-.017	-.017	-.032	-.029	1.183	1.205	1.135	1.155
NLL	.381	.380	.461	.404	.016	.025	.031	.039	1.376	1.357	1.501	1.403
PHT	.367	.348	.349	.348	.005	-.010	-.009	-.025	1.308	1.295	1.281	1.300
NHT	.336	.342	.341	.359	.005	.016	.006	.043	1.255	1.261	1.243	1.293
PLT	.348	.337	.356	.355	-.006	-.009	-.014	-.016	1.273	1.272	1.266	1.299
NLT	.351	.356	.338	.342	.013	.017	.012	.018	1.290	1.292	1.269	1.253

Note. See Table 1 for the condition names. Abs_Bias_S = Absolute bias under the small non-invariance magnitude conditions; Abs_Bias_M = Absolute bias under the medium non-invariance magnitude conditions; Bias_S = Bias under the small non-invariance magnitude conditions; Bias_M = Bias under the medium non-invariance magnitude conditions; RMSE _S = RMSE under the small non-invariance magnitude conditions; RMSE _M = RMSE under the medium non-invariance magnitude conditions.

Appendix

Measurement Invariance Testing in Multiple Group CFAs

A measurement invariance test is simultaneously conducted for multiple groups by fitting nested models across groups. Stepwise procedures are commonly used, beginning with the least restrictive model (e.g., configural invariance) and increasing the restrictiveness of constraints (e.g., scalar invariance). By subsequently constraining the factor structure, factor loadings, and intercepts/thresholds across all items, measurement invariance is tested between two nested models by using a chi-square difference test. In M*plus*, the DIFTEST command produces the results of a chi-square test for difference testing for estimators including ULSMV. If statistical significance is found in a chi-square difference test, measurement non-invariance exists in the parameters constrained in the more restrictive model between subgroups.

Study 1 Convergence and Model Fit

When the negative medium magnitude of metric non-invariance (-0.6) was manipulated for five items with relatively low loadings in absolute value (NLL conditions), nonconvergence rates of 10% and 6% were found for configural and metric invariance respectively (Table 2A). In terms of model fit, all converged models showed RMSEA values $\leq .06$. When the cutoff CFI $<$ .95 was used the proportions of models with poor fit ranged from .00 to .03, and the CFI values from poor fitting models were very close to the criterion; the minimum CFI was .93, and both the maximum and average were .94. The proportions were slightly higher under the CI conditions than the other conditions (Table 3A).

Regarding the nonconvergence rates of the NLL conditions, by closely looking at the manipulation results it was found that the manipulation led three negatively keyed items to be more negative (e.g., the loading of -0.5 became -1.1) and two positively keyed items to not load on the corresponding factor (the loading of 0.6 became 0). Especially in configural invariance, non-loading items may potentially lead to different factor structures between the reference and focal groups, resulting in estimation difficulties.

Interestingly, all models under NLL conditions were converged when ten items were manipulated for measurement non-invariance. We postulated that this may be related to forced-choice formats' interdependent nature in estimation process. That is, more appreciable systematic changes in more items (two items per factor, as a total 10 items) than in the five item conditions (one item per factor) detected through the interdependent estimation process seems to help the models to be converged. As discussed later in the results section, it was found that absolute bias from NLL under the five-item conditions were larger than the ten-item conditions especially when the magnitude of the measurement non-invariance was relatively larger. This finding also indicates that the manipulation of a smaller number of items produced more biased estimates than that with larger number of items and more adverse impact was found possibly due to the difficulty in the computation process caused by relatively less systematic changes that the interdependent estimation process can detect. It should be noted that this was based on the authors' speculation, and thus requires a more in-depth investigation.

Study 2 t-test and ANOVA Results

Type I error rates and detection of measurement non-invariance. To examine factors affecting Type I error rates and the detection of measurement non-invariance, results from the recommended cutoffs of Δ CFI and Δ NCI were utilized. Due to the small sample sizes for detection (n = 56) and Type I error rates (n = 24), individual t-tests were employed for each factor with an adjusted p value of .0125 to control for multiplicity, instead of ANOVA. Effect size interpretations were based on Cohen (1988). A repeated measures t-test revealed that Δ CFI performed significantly better than Δ NCI; Δ CFI cutoffs showed smaller Type I error rates by .03 and higher detection by .04. The effect size was large for Type I error rates (*d* = 1.5) and medium for detection (*d* = 0.5).

Among the factors affecting Type I error rates, the direction of measurement non-invariance was a significant factor. The t-test results found that when measurement non-invariance was in the negative direction, significantly higher Type I error rates were detected by .03 (*d* = 1.5) with the use of Δ NCI cutoffs. As expected, the magnitude of measurement non-invariance significantly affected non-invariance detection for both Δ CFI and Δ NCI; detection levels were higher by .44 (Δ CFI, *d* = 1.7) and .43 (Δ NCI, *d* = 1.6) when the medium magnitude of measurement non-invariance was manipulated. Combining the results from the t-tests, using Δ CFI can be recommended to avoid false detection of measurement non-invariance especially when the negative direction of non-invariance was expected.

Bias and RMSE. In terms of the factors affecting absolute bias, ANOVA[1] results indicated that item characteristics (high vs. low loadings/thresholds), directions of non-invariance (positive vs. negative), and types of non-invariance (metric vs. scalar) were found as significant factors, although the effect sizes were quite small ($\eta^2 = .001, \eta^2 = .002, \eta^2 = .003$, respectively). The findings from absolute bias showed that the impact of failure in the detection of metric non-invariance was more adverse than that of scalar non-invariance by .02. Additionally, the negative direction of non-invariance revealed slightly larger absolute bias by .02.

The direction of non-invariance showed more meaningful results when bias and RMSE were examined. The direction non-invariance systematically affected bias; negative non-invariance increased estimated trait scores on average by .02, whereas positive non-invariance

[1] For bias and RMSE ANOVAs were conducted because the sample size was large (n = 3,200).

led to a decrease in trait scores by .01. The ANOVA results showed the direction of non-invariance was a significant factor along with a medium effect size ($\eta^2 = .08$). The same pattern was detected with RMSE; the direction of non-invariance was a significant factor with a small effect size ($\eta^2 = .02$). Combining the ANOVA results for bias and RMSE, failure in the detection of differential loadings between focal and reference groups can threaten validity in the use of scores from a forced-choice format. Negative changes in item parameters potentially distort scores more than positive changes of the same magnitude.