Data Generation to illustrate clinical applications of ESEM

© Alexandre J.S. Morin, Philip D. Parker, & Herbert W. Marsh,
University of Western Sydney

To illustrate ESEM analyses relevant to clinical applications, we simulated a multivariate normal data set. This data set is made available to readers willing to try their hands at ESEM at https://github.com/pdparker/ESEM, together will all of the input codes used in this specific illustration. This data set includes six items (X_1 through X_6) serving as indicators of two correlated factors (Factor 1 and Factor 2), with the first three having their primary loadings on the first factor and the last three having their primary loadings on the second factor, and with two items presenting significant cross loadings on the other factor. For purposes of illustration, we label these factors Anxiety (Factor 1) and Depression (Factor 2), two correlated/comorbid clinical states which can be measured by indicators/symptoms that can realistically be expected to present significant, and even reasonably large, cross-loadings (for instance, levels on the depressive symptoms of "psychomotor agitation" or "insomnia" can be expected to be also elevated in anxious individuals). For further illustrative purposes, lets imagine that these data were collected as part of a clinical *Pre-Test Post-Test* design with randomized experimental and control groups. In this context, the first set of items (X_1 through X_6) will be referred to as *Pre-Test* data. We also simulated a second set of items (Y_1 through Y_6), referred to as *Post-Test* data, designed to represent a second measurement point for the *X* items and simulated to have similar properties to the *Pre-Test* data. Two subgroups were simulated to reflect an experimental and a control group. In line with clinical trials, these subgroups were simulated based on the assumption of random assignments (with *Pre-Test* data showing fully equivalent properties between the groups) and moderately-small sample sizes of $n = 150$ for each group ($n = 300$ in total). Further, consistent with common observations in the context of clinical trials, the control group was simulated to show a small decrease of depressive symptoms over time but no change in anxiety levels. The experimental group was simulated as showing an substantial decrease in depressive symptoms over time, and with gender-differentiated effects regarding the response to treatment for symptoms of anxiety (see below), consistent with significant construct-specific intervention effects. We simulated the data to show increased levels of measurement errors associated with the last two items of the depression factor in the experimental group, something that reflects the occurrence of some disturbance (e.g., noise, fire alarm) toward the end of the testing session in one of the groups.

To add some complexity, we further simulated slightly different models for subgroups reflecting males (n = 150) versus females (n = 150) participants. Consistent with random assignment, half (n = 75) of the experimental and control groups was simulated as males, while the other half was simulated as including females. We simulated the data as showing one item with higher intercept among females, consistent with well-documented gender differences on specific indicators of internalized disorders (e.g. self-esteem, appetite loss, etc.). Although gender differences in mean levels of depression and anxiety (favouring females) are well documented, here we simulated data consistent with a clinical trial where all participants are selected to be fully comparable at baseline. For this reason, we simulated the data to show no gender differences in mean-levels of Depression or Anxiety at *Pre-Test*. We also simulated the data to show differential response to treatment showing that the previously described reduction in Depression was common to males and females. However, females also showed a significant decrease in anxiety at *Post-Test*, whereas males showed an increase in anxiety levels at *Post-Test*, showing a deleterious effect of treatment.

The population generating model is presented in Figure 1. All analyses were conducted with Mplus 7.0 (Muthén & Muthén, 2011), using the default maximum likelihood (ML) estimator—although other options might be preferable for other situations (e.g., multilevel, ordered-categorical, or missing data) that can be implemented in conjunction with ESEM. For pedagogical purposes, we now move to analyses of this simulated data set. The dataset, as well as input codes, are available at https://github.com/pdparker/ESEM.

Comparison of ESEM and CFA models

ESEM investigation should always begin by a test of whether the a priori factor model fits the data well and whether the ESEM model provides a better fit to the data than a traditional and more parsimonious ICM-CFA model. Thus, CFA and ESEM models were first estimated separately on *Pre-Test* and *Post-Test* data. For purposes of illustrations, we focus here on target rotation, which is the more complex rotation to specify, and also the most flexible in terms of allowing the estimation process to be guided a priori by theory. The goodness of fit results from these models are reported at the top of Table 1. These results indicate that the ICM-CFA model provides a suboptimal level of fit to the data at both *Pre-Test* and *Post-Test* (significant χ^2; TLI < .95; RMSEA > .08). Conversely, the ESEM model provides an almost perfect fit to the data at both *Pre-Test* and *Post-Test* (non-significant χ^2; CFI and TLI > .95; RMSEA < .06).

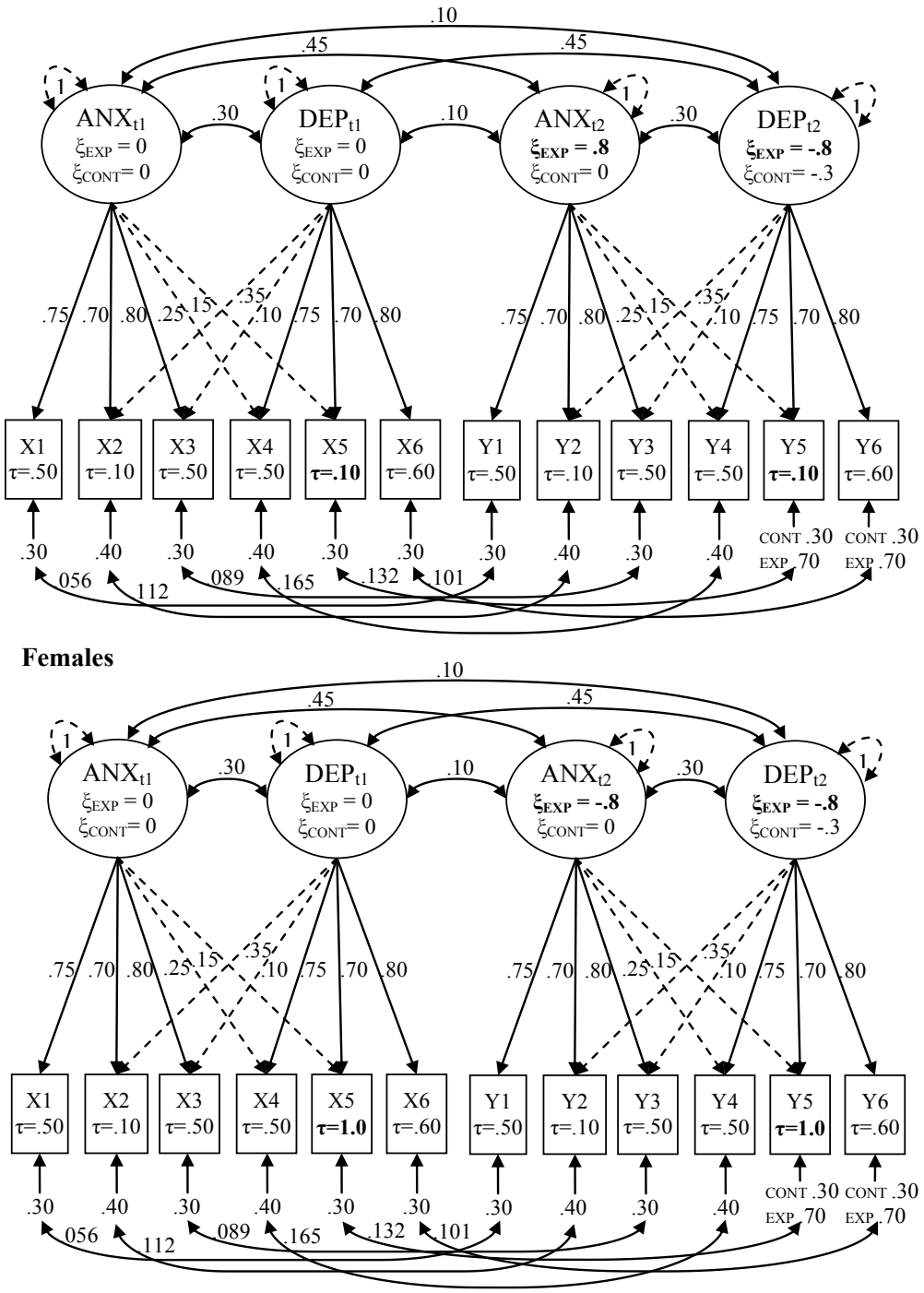

Figure 1. Graphical representation of the population generating model.

Note. Ovals represent latent factors and squares represent observed variables; full unidirectional arrows linking ovals and squares represent the main factor loadings; dotted unidirectional arrows linking ovals and squares represent the cross-loadings; full unidirectional arrows placed under the squares represent the item uniquenesses; bidirectional full arrows linking the ovals represent factor covariances/correlations; bidirectional full arrows linking the squares represent the longitudinal correlated uniquenesses; bidirectional dashed arrows connecting a single oval represent factor variances; τ represents items intercepts; ξ represents the latent factor means; the subscripts EXP and CONT indicate that a parameter varies across both simulated subgroups; the labels ANX and DEP refer to the factors designed to respectively reflect anxiety and depression; the t1 and t2 respectively refer to the *Pre-Test* and *Post-Test*.

It is also instructive to compare parameter estimates based on the ICM-CFA and ESEM solutions, as reported in Table 2. The main difference – in addition to the ESEM cross-loadings – is that the ICM-CFA model results in slightly higher estimates of the factor correlations due to the unrealistic assumption of 0 cross loadings. Although this difference is not as pronounced in this specific application based on 6-items, 2-factors (with differences in correlations approaching .100 between ESEM and CFA), and low cross loadings, it remains important as illustrated by the important increases in fit indices. Similarly, although the estimated ESEM correlations (r = .525-.579) remains higher than the one used in the population generating model (r = .300) this is also due to the issue of rotational indeterminacy. Thus, if our objective had been to maximally decrease the factor correlation, we would have explored alternative rotation procedures, which we recommend should always be done in ESEM. For instance, using Geomin rotation with an epsilon value of .5, as recommended by Marsh et al. (2009) to maximally deflate factor correlations, the estimated factor correlation (r = .350-.389) becomes much closer to the population value.

Measurement invariance: The multiple group approach

To illustrate multigroup ESEM approach to measurement invariance, we tested whether the *Pre-Test* measurement model was invariant across the treatment and control groups. This is a critical, though often untested, assumption of randomized control trials where the groups are expected, and often assumed, to be fully equivalent at the *Pre-Test*. To illustrate how these tests can further be extended to multiple grouping variables, we also conducted tests of measurement invariance across combinations of treatment status and gender to ensure that the measurement model was not only equivalent across treatment groups, but also across genders, within and across treatment groups. Indeed, when the multiple groups reflect a factorial combination of grouping variables (e.g., 2 genders X 2 treatment groups= 4 groups), it is possible to discern non-invariance due to the interaction of the two grouping variables as well as the "main" effects of each grouping variable considered separately. The interested readers may use the provided data set to further their ESEM capabilities by developing their own inputs codes to further test the invariance of the measurement model across genders at *Pre-Test*, as well as across treatment status, gender, and their combination at *Post-Test*. We also provide a much more extensive example later one (invariance across 2 time points X 2 genders X 2 treatments) from which it is possible to deduce this more simple application.

Table 1. Results from the total group and multiple group cross sectional models

Model	χ^2	df	RMSEA	RMSEA 95% CI	CFI	TLI
Main models						
Pre-Test CFA (CFA-Pre.inp)	32.990*	8	.102	.067-.139	.969	.942
Pre-Test ESEM (ESEM-Pre.inp)	5.499	4	.035	.000-.100	.998	.993
Post-Test CFA (CFA-Post.inp)	37.827*	8	.111	.077-.148	.964	.933
Post-Test ESEM (ESEM-Post.inp)	2.403	4	.000	.000-.069	1.000	1.007
Multiple group (Treatment) invariance (Pre-Test)						
Model 1: Configural (MG-inv-M1.inp)	9.640	8	.037	0-.107	.998	.992
Model 2: Load. (MG-inv-M2.inp)	10.448	16	.000	0-.044	1	1.013
Model 3: Load., uniq. (MG-inv-M3.inp)	17.650	22	.000	0-.051	1	1.007
Model 4: Load., FVC. (MG-inv-M4.inp)	13.077	19	.000	0-.043	1	1.012
Model 5: Load., int. (MG-inv-M5.inp)	12.195	20	.000	0-.028	1	1.015
Model 6: Load., uniq., FVC. (MG-inv-M6.inp)	20.649	25	.000	0-.051	1	1.007
Model 7: Load., int., uniq. (MG-inv-M7.inp)	19.400	26	.000	0-.041	1	1.009
Model 8: Load., int., FVC. (MG-inv-M8.inp)	14.821	23	.000	0-.029	1	1.013
Model 9: Load., int., uniq., FVC. (MG-inv-M9.inp)	22.393	29	.000	0-.041	1	1.009
Model 10: Load., int., FMeans. (MG-inv-M10.inp)	13.393	22	.000	0-.024	1	1.015
Model 11: Load., int., uniq., FMeans. (MG-inv-M11.inp)	20.598	28	.000	0-.037	1	1.010
Model 12: Load., int., FVC., FMeans. (MG-inv-M12.inp)	16.025	25	.000	0-.025	1	1.013
Model 13: Load., int., uniq., FVC., FMeans. (MG-inv-M13.inp)	23.600	31	.000	0-.038	1	1.009
Multiple group (Gender X Treatment) invariance (Pre-Test)						
Model 1: Configural (MG-inv-M1.inp)	14.034	16	.000	.000-.094	1	1.008
Model 2: Load. (MG-inv-M2.inp)	35.919	40	.000	.000-.067	1	1.007
Model 3: Load., uniq. (MG-inv-M3.inp)	55.459	58	.000	.000-.066	1	1.003
Model 4: Load., FVC. (MG-inv-M4.inp)	45.648	49	.000	.000-.066	1	1.005
Model 5: Load., int. (MG-inv-M5.inp)	120.686	52	.133	.102-.164	.923	.911
Model 5p: Load., p.int. (MG-inv-M5p.inp)	42.972	51	.000	.000-.052	1	1.011
Model 6: Load., uniq., FVC. (MG-inv-M6.inp)	64.715	67	.000	.000-.064	1	1.002
Model 7: Load., int., uniq. (MG-inv-M7.inp)	137.924	70	.114	.086-.142	.924	.935
Model 7p: Load., p.int., uniq. (MG-inv-M7p.inp)	62.248	69	.000	.000-.054	1	1.007
Model 8: Load., int., FVC. (MG-inv-M8.inp)	127.958	61	.121	.091-.150	.925	.926
Model 8p: Load., p.int., FVC. (MG-inv-M8p.inp)	52.665	60	.000	.000-.053	1	1.008
Model 9: Load., int., uniq., FVC. (MG-inv-M9.inp)	145.340	79	.106	.078-.133	.926	.944
Model 9p: Load., p.int., uniq., FVC. (MG-inv-M9p.inp)	71.47	78	.000	.000-.053	1	1.006
Model 10: Load., int., FMeans. (MG-inv-M10.inp)	188.495	58	.173	.146-.201	.854	.849
Model 10p: Load., p.int., FMeans. (MG-inv-M10p.inp)	47.562	57	.000	.000-.047	1	1.011
Model 11: Load., int., uniq., FMeans. (MG-inv-M11.inp)	208.941	76	.153	.128-.178	.851	.882
Model 11p: Load., p.int., uniq., FMeans. (MG-inv-M11p.inp)	66.871	75	.000	.000-.050	1	1.007
Model 12: Load., int., FVC., FMeans. (MG-inv-M12.inp)	198.571	67	.162	.136-.188	.853	.868
Model 12p: Load., p.int., FVC., FMeans. (MG-inv-M12p.inp)	57.323	66	.000	.000-.049	1	1.009
Model 13: Load., int., uniq., FVC., FMeans. (MG-inv-M13.inp)	218.553	85	.145	.121-.169	.850	.894
Model 13p: Load., p.int., uniq., FVC., FMeans. (MG-inv-M13p.inp)	76.150	84	.000	.000-.050	1	1.006
MIMIC models (Gender)						
Post-Test MIMIC null (MIMIC-null-Post.inp)	104.753*	10	.178	.148-.209	0.898	0.786
Post -Test MIMIC saturated (MIMIC-satur-Post.inp)	2.009	4	.000	.000-.062	1.000	1.011
Post -Test MIMIC invariant intercept (MIMIC-base-Post.inp)	32.135*	8	.100	.066-.138	.974	.932
Post -Test MIMIC partially invariant intercept (MIMIC-DIF-Post.inp)	3.593	7	.000	.000-.043	1.000	1.011

Note. Names of the input file in the supplementary materials are reported in parentheses; *: $p < .05$; CFA: Confirmatory factor analysis; ESEM: Exploratory Structural Equation Modeling; χ^2: Chi square test of model fit; df: degrees of freedom; RMSEA: Root Mean Square Error of Approximation; RMSEA 95% CI: 95% confidence interval of the RMSEA; TLI: Tucker-Lewis Index; CFI: Comparative Fit Index; Load: Loadings invariance; Uniq: Uniquenesses invariance; FVC: Factor variance-covariance invariance; Int: Intercepts invariance; FMeans: Factor means invariance.

Table 2. Standardized parameters from the CFA and ESEM models based on cross sectional data.

	CFA			ESEM		
Item	**Anxiety**	**Depression**	**Uniq.**	**Anxiety**	**Depression**	**Uniq.**
Time 1		(CFA-Pre.inp)			(ESEM-Pre.inp)	
X1	.735		.459	.824	-.105	.410
X2	.871		.241	.726	.191	.277
X3	.827		.316	.874	-.035	.270
X4		.830	.311	.170	.681	.374
X5		.711	.494	-.026	.745	.467
X6		.719	.483	-.083	.819	.402
Correlations	.659			.579		
Time 2		(CFA-Post.inp)			(ESEM-Post.inp)	
Y1	.826		.318	.906	-.111	.272
Y2	.859		.261	.755	.174	.262
Y3	.874		.237	.875	-.003	.236
Y4		.828	.314	.191	.647	.415
Y5		.657	.568	-.134	.816	.431
Y6		.619	.617	.012	.644	.577
Correlations	.610			.525		

Note. Names of the input file in the supplementary materials are reported in parentheses; All coefficients significant at the .05 level; CFA: Confirmatory factor analysis; ESEM: Exploratory Structural Equation Modeling; F1: standardized loadings on the first factor; F2: standardized loadings on the second factor; Uniq.: standardized uniquenesses.

The 13-model taxonomy of invariance tests is an important contribution of research in the area of ESEM, although the same logic applies to CFA models. However, coding and running all models by hand can be time consuming and, for complex models, prone to error. As such, we are developing a set of function in R (R core development team 2013) to automate the construction of Mplus scripts with only a few lines of code. The latest version of these functions, simulated datasets, associated documentation, and example workflow can be found (https://github.com/pdparker/ESEM). Currently this function is available for cross-sectional data only. When paired with the MplusAutomation package (Hallquist & Wiley, 2013) a few lines of R code can write, run, and display in a table the results all 13 models (see Figure 2, based on exampleScript_ESEMS.R from https://github.com/pdparker/ESEM). Functions are available for both target and Geomin rotation. If one is only interested in invariance by inspecting fit criteria then it does not matter which rotation is used as fit is independent of rotation. Here we use the Geomin function as it requires fewer function arguments to specify and runs quicker than target rotation.

Treatment group invariance

Mplus Summary Table

	Title	ChiSqM_Value	ChiSqM_DF	CFI	TLI	RMSEA_Estimate	RMSEA_90CI_LB	RMSEA_90CI_UB
1	ESEM Model 01;	9.640	8	0.998	0.992	0.037	0	0.107
6	ESEM Model 02;	10.448	16	1.000	1.013	0.000	0	0.044
7	ESEM Model 03;	17.650	22	1.000	1.007	0.000	0	0.051
8	ESEM Model 04;	13.077	19	1.000	1.012	0.000	0	0.043
9	ESEM Model 05;	12.195	20	1.000	1.015	0.000	0	0.028
10	ESEM Model 06;	20.649	25	1.000	1.007	0.000	0	0.051
11	ESEM Model 07;	19.400	26	1.000	1.009	0.000	0	0.041
12	ESEM Model 08;	14.821	23	1.000	1.013	0.000	0	0.029
13	ESEM Model 09;	22.399	29	1.000	1.009	0.000	0	0.041
2	ESEM Model 10;	13.393	22	1.000	1.015	0.000	0	0.024
3	ESEM Model 11;	20.598	28	1.000	1.010	0.000	0	0.037
4	ESEM Model 12;	16.025	25	1.000	1.013	0.000	0	0.025
5	ESEM Model 13;	23.600	31	1.000	1.009	0.000	0	0.038

Treatment group by gender invariance

Mplus Summary Table

	Title	ChiSqM_Value	ChiSqM_DF	CFI	TLI	RMSEA_Estimate	RMSEA_90CI_LB	RMSEA_90CI_UB
1	ESEM Model 01;	14.034	16	1.000	1.008	0.000	0.000	0.094
6	ESEM Model 02;	35.919	40	1.000	1.007	0.000	0.000	0.067
7	ESEM Model 03;	55.459	58	1.000	1.003	0.000	0.000	0.066
8	ESEM Model 04;	45.648	49	1.000	1.005	0.000	0.000	0.066
9	ESEM Model 05;	120.686	52	0.923	0.911	0.133	0.102	0.164
10	ESEM Model 06;	64.715	67	1.000	1.002	0.000	0.000	0.064
11	ESEM Model 07;	137.924	70	0.924	0.935	0.114	0.086	0.142
12	ESEM Model 08;	127.958	61	0.925	0.926	0.121	0.091	0.150
13	ESEM Model 09;	145.340	79	0.926	0.944	0.106	0.078	0.133
2	ESEM Model 10;	188.495	58	0.854	0.849	0.173	0.146	0.201
3	ESEM Model 11;	208.941	76	0.851	0.882	0.153	0.128	0.178
4	ESEM Model 12;	198.571	67	0.853	0.868	0.162	0.136	0.188
5	ESEM Model 13;	218.553	85	0.850	0.894	0.145	0.121	0.169

Figure 2. Output using MplusAutomation and ESEM helper functions.

As can be seen from Table 1 and Figure 2, the results show strong support for the full invariance of the measurement model, as well as the latent variance-covariances and latent means, across treatment group at the pretest. For the treatment group by gender interaction there was strong evidence for factor loading invariance but not intercept invariance. This suggests that comparing groups on models based on the covariance matrix is justified, however, comparing latent means is potentially dubious and researchers should consider partial invariance of the intercepts. Freeing the intercept in model 5 (and subsequently for models 7-13) for variable X5 across gender but not treatment group provided a substantial improvement in fit. This would allow researchers to tentatively compare latent means across gender while noting the need to interpret results in light of the partial invariance. This process can easily be extended to post-test and is left to the reader as an exercise. The remaining steps of this taxonomy of invariance test confirmed that the *Pre-Test* uniquenesses, latent variance-covariances and latent means were also fully invariant across combinations of gender and treatment groups.

Measurement invariance: The MIMIC approach.

In a MIMIC model latent variables are regressed on observed predictors. When the latent variables have multiple indicators, the MIMIC model can be extended to test potential non-invariance of item intercepts, that is, DIF (technically, monotonic DIF). In applied clinical research based on often modest sample sizes, the MIMIC model has the advantage of being more parsimonious than the multiple group approach to measurement invariance, and can accommodate multiple predictors, continuous predictors, as well as interactions among predictors. However, while the MIMIC model is able to test DIF, and can be extended to tests of the invariance of factor loadings (non-monotonic DIF, for details see Barendse, Oort, & Garst, 2010; Barendse, Oort, Werner et al., 2012; Woods & Grimm, 2011), it cannot be used to test for the invariance of items' uniquenesses. As previously described, monotonic DIF is evaluated by comparing three models. The null effect MIMIC posits that the predictor variables have no effect on the latent factors or items intercepts. The saturated MIMIC has paths from each predictor variable to all item intercepts, but not the latent factors. The invariant intercept MIMIC has freely estimated paths from the predictors to the latent factors, but not the item intercepts. The comparison of Model 1 with Models 2 and 3 tests whether there are any effects of the predictors, the comparison of model 1 and model 3 tests whether the predictors have an effect on the latent variables, while the comparison of Model 2 and Model 3 tests whether the effects of the predictor variables on individual items can be fully explained in terms of effects on the latent factors. If Model 2 fits substantially better than

Model 3, then there is evidence of monotonic DIF (i.e., non-invariance of intercepts). In this case, it might be appropriate to pursue partially invariant models in which the invariance constraint is relaxed for some intercepts.

Here, we focus here on the use of the MIMIC model as a way to investigate monotonic DIF as a function of gender (since this data was simulated to show DIF on one item as well as latent mean differences) based on *Post-Test* data but we invite the readers to try their hands at the MIMIC approach based on *Pre-Test* data and treatment group. A more complex application of this MIMIC model is presented in the longitudinal section. For our simulated data (see Table 1), the MIMIC null effect model, in which the grouping variable is posited to be unrelated to the ESEM factors or the items, failed to provide an acceptable fit to the data (significant χ^2; CFI and TLI < .95; RMSEA > .08). This suggests that at least some effects of the predictor variable should be expected. Indeed, the saturated MIMIC model did provide a satisfactory fit to the data (non-significant χ^2; CFI and TLI > .95; RMSEA < .06) and a substantial improvement over the null effect model. The third (intercept invariant) MIMIC model (i.e., in which the grouping variable is only allowed to predict the latent factor scores but not the items) also failed to provide an acceptable fit (significant χ^2; TLI < .95; RMSEA > .08), suggesting DIF. Examination of the modification indices associated with this model indicates that DIF was mainly associated with item 5. Allowing for direct effects of the predictor on item 5, in addition to its effects on the ESEM factors, results in a satisfactory fit to the data (non-significant χ^2; CFI and TLI > .95; RMSEA < .06) and in a fit that is comparable to the fit of the saturated MIMIC model [$\Delta\chi^2$ (df) = 1.584 (3), $p > .05$; ΔTLI, ΔCFI and ΔRMSEA = 0]. Detailed results from this model reveal that participants' levels on the anxiety factor ($\hat{\beta}$ = -.313, $p < .001$) tend to be higher among males, that scores on item 5 tended to be higher among females ($\hat{\beta}$ = .354, $p < .001$), while scores on the depression factors show no significant effect of gender ($\hat{\beta}$ = -.083, *ns*).

Measurement invariance: The longitudinal approach

Essentially the same logic and the same taxonomy of models can be used to test the invariance of parameters across multiple occasions for a single group. One distinctive feature of longitudinal analyses is that they should normally include correlated uniquenesses (CUs) between responses to the same item on different occasions (see Jöreskog & Sörbom, 1977; Marsh, 2007; Marsh & Hau, 1996, 2004). When the same items are used on multiple occasions, the uniqueness component associated with each item from one occasion is

typically positively correlated with the uniqueness component associated with the same item on another occasion. Failure to include these CUs generally results in biased parameter estimates. In particular, test-retest correlations among matching latent factors are systematically inflated, which can then systematically bias other parameter estimates and may even result in improper solutions such as a non-positive definite factor variance-covariance matrix or estimated test-retest correlations that exceed 1.0 (e.g., Marsh, Martin, & Hau, 2006; Marsh, Martin, & Debus, 2001). Interestingly, the inclusion of CUs is another option within ESEM that was not typically possible in traditional EFAs.

We first compared longitudinal measurement models excluding and including correlated uniquenesses between matching items (i.e., X_1 with Y_1, X_2 with Y_2, etc) (see first four models in Table 3). The comparison of these models clearly showed that the models excluding the CUs did not fit the data well (CFI and TLI < .95 and even .90; RMSEA > .10). In contrast, the CFA with CUs was acceptable according to some criteria (CFI, TLI > .95; RMSEA =.070)—even though the fit was substantially improved for the ESEM model with CUs (non-significant χ^2, CFI, TLI > .98; RMSEA =.033). However, inspection of the factor correlations based on these four models, which are reported in Table 6, reveals that CFA factor correlations within each time point were systematically slightly higher (in line with what was observed with the cross sectional models, a difference that is again more pronounced if we use Geomin rotation) when compared to the ESEM factor correlations. Of particular relevance, the test-retest correlations are systematically higher for models that do not contain CUs (and inflated relative to the known factor correlations for our simulated data) than for models that do. On this basis, we evaluated longitudinal invariance in relation to ESEM models that included CUs.