The results from the 13 steps of longitudinal measurement invariance tests, ignoring grouping variables, are reported in Table 3 (Models L1 to L13). As the results from the longitudinal invariance tests closely parallel those from the multiple group invariance tests, they will not be described in detail. These results support the complete longitudinal invariance of the factor loadings, of the factor variance-covariance matrix, and of the items' intercepts. Interestingly, the results also provide some support for the longitudinal invariance of the items' uniquenesses, this support remains marginal across comparisons: Model 2 vs. Model 3 (ΔRMSEA = +.013; ΔCFI = -.007; ΔTLI = -.008); Model 4 vs. Model 6 (ΔRMSEA = +.011; ΔCFI = -.008; ΔTLI = -.009); Model 5 vs. Model 7 (ΔRMSEA = +.013; ΔCFI = -.007; ΔTLI = -.009); Model 8 vs. Model 9 (ΔRMSEA = +.013; ΔCFI = -.007; ΔTLI = -.008); Model 10 vs. Model 11 (ΔRMSEA = +.002; ΔCFI = -.007; ΔTLI = -.003); Model 12 vs. Model 13 (ΔRMSEA = +.003; ΔCFI = -.008; ΔTLI = -.005). In practice, observing this pattern of results would suggest the need to probe the results for possible non-invariance of the items uniquenesses. Examination of the modifications indices associated with the models in which the uniquenesses are constrained to be invariant would suggest that the invariance constraint might be relaxed for item 6, and maybe also item 5. However, relaxing these constraints would result in generally negligible differences in the size of the estimated uniquenesses across time waves, which is consistent with the population generating model in which the non-invariance of the items's uniquenesses was simulated to be a function of an interaction effect between the treatment group and the time waves (see next section). The preliminary cross-sectional analyses reported in the previous section would already have alerted the researcher to the need to consider potential interaction effects between time and treatment group in the examination of possible non-invariance for the items' uniquenesses.

Finally, the last four models all converged in rejecting the longitudinal invariance of the factor means. Exploration of results based on models including invariance of the items' intercepts revealed that, when *Pre-Test* latent means are constrained to be zero, *Post-Test* latent means vary between -.051 and -.056 (non-significantly different from 0) on the Anxiety factor and on -.546 and -.563 on the Depression factor, close to the population values of 0 for Anxiety (remember that *Post-Test* levels of anxiety are diametrically opposed for male and females, and average 0 when gender groups are combined) and -.550 for Depression (remember that *Post-Test* levels of depression are substantially higher in the Treatment group and average –.550 on the total sample).

Illustrating a Clinical Trial through Multiple-Groups Longitudinal Models

It is easy to extend the preceding model to the investigation of a clinical trial. Here, two groups are measured at two points in time (*Pre-Test* and *Post-Test*) and treatment effects would be reflect in an increase in latent means over time that would be substantially higher in the treatment group than in the experimental group. This approach is highly similar to the repeated-measures ANOVA approach typically used to evaluate treatment effects, but based on fully latent variable models. The advantage of models based on fully latent variables, versus traditional ANOVAs based on scale scores, is that latent variables models provides a way to test that the variables do indeed reflect the same constructs across all possible combinations of groups and time points, as well as a way to control for measurement errors. As previously noted, all time or group-based comparisons rest on the strong assumption that the constructs are measured in the same way across groups, time, and group by time interactions, an assumption that can be directly tested in latent variable models. In practice, these assumptions are explored through tests of measurement invariance conducted simultaneously across time points and treatment groups (a 2 time points X 2 groups measurement invariance model). The results from these analyses are reported in Table 3.

The results from the 13 steps of longitudinal measurement invariance tests across treatment groups, ignoring gender, are reported in Table 3 (see Models LT1 to LT13). These results support the complete invariance of the factor loadings, the factor variance-covariance matrix, and the items' intercepts across time waves and treatment groups. However, the ΔCFI results point to the non-invariance of the items' uniquenesses (ΔCFI = -.010 to -.013 across comparisons). Probing for this non-invariance led to the conclusions that the items uniquenesses where not invariant across time waves in the experimental group for the last two items which showed a substantially higher level of measurement error. For this reason, models including the partial invariance of the items' uniquenesses were also estimated (marked by p in Table 3), which confirmed the complete invariance of the uniquenesses associated with the remaining items. As in the previous illustration, the last four models converged in rejecting the longitudinal invariance of the factor means. Exploration of results based on models including invariance of the items' intercepts revealed that, when *Pre-Test* latent means are constrained to be zero in the control group: (a) in the Experimental group, *Pre-Test* latent means vary between .037 and .042 on the Anxiety factor and between .130 and .152 on the Depression factor (consistent with the equivalence of the groups); (b) in the Control group, *Post-Test* latent means vary between -.005 and -.006 (non-significantly different from 0) on the Anxiety factor and on -.249 and -.275 on the Depression factor; (c) in

the Experimental group, *Post-Test* latent means vary between -.057 and -.067 (non-significantly different from 0, consistent with non-modeled gender-differentiated effects) on the Anxiety factor and between -.738 and -.806 on the depression factor. Once again, these values are in line with population values.

Illustrating a Clinical Trial through Multiple-Groups Longitudinal Models: Gender-Differentiated effects.

It is often the case that researchers are not only interested in investigating the effects of a grouping variable or a treatment program, but also whether different subgroups show different reactivity to these effects. Here, we thus consider an additional grouping variable, that we simulated so as to represent gender – an important covariate to consider in studies of anxiety and depression. Extending the previous analyses to incorporate gender as an additional grouping variable is easy. To illustrate one possible way of doing this, we replicated the previous tests of invariance across interactions of 2 time waves X 2 treatment conditions X 2 gender groups. These tests allow for the investigations of differential treatment effects for males and females, tests which also rely on the assumption of measurement invariance. The results from the 13 steps of measurement invariance tests simultaneously taking into account gender, treatment groups, and time waves, are reported in Table 3 (see model LTG1 to LTG13). As the results from the longitudinal invariance tests closely parallel those from the multiple group invariance tests, they will not be described in detail. These results support the complete longitudinal invariance of the factor loadings and of the factor variance-covariance matrix, but clearly show that the items uniqueness, items' intercepts, and latent means were not invariant. Regarding the items' uniquenesses, the results show that the uniquenesses associated with the last two items were slightly higher at *Post-Test* in the experimental group, an effect that is consistent across gender. Regarding the items' intercepts, the results show that the intercept associated with item 5 (X5 and Y5) was higher among females than males, an effect that was consistent across time waves and experimental groups. We thus pursued partial invariance models in which invariance constraints were relaxed for these items' intercepts (model pi in Table 3), uniquenesses (models pu in Table 3) and combinations of both (models pipu in Table 3). Latent means differences were explored across models including partial invariance of the items' intercepts, and are represented in Table 5. Once again, these results are in line with those from the population model, confirming the equivalence of all groups at *Pre-Test* as well as a significant decrease in levels of Depression at Post -Test in all groups, but a more pronounced decrease in the experimental group. Similarly, these results show a gender differentiated

effect on levels of anxiety showing a significant increase in Anxiety levels for males at *Post-Test* compared to their *Pre-Test* levels, and a significant decrease in levels Anxiety levels for females compared to *Pre-Test* levels.

It must be noted that, although it was possible to estimate the preceding gender X treatment X time waves models in this specific application based on a well-defined simulated dataset, the sample sizes used here (with $n = 75$ per gender X treatment group) – selected to reflect common sample sizes used in the context of clinical studies – is likely to be too low to allow for the same tests to be conducted based on real datasets. Indeed, real data often pose multiple additional challenges during the estimation process linked to multiple sources of measurement errors, violations of distributional assumptions, etc. For these contexts, we propose the previously described MIMIC model as a way to probe for gender-based difference in reactivity to treatment without having to rely on complex multiple groups models such as those described in the previous section. Furthermore, an additional advantage of the MIMIC approach is the ability to consider continuous predictors (i.e. such as age) and interactions among predictors. To maintain the analogy with ANOVAs, these models are akin to ANCOVAS where the effect of one predictor can be controlled for in the estimation of group-based differences. However, the MIMIC approach goes further than that in allowing for an investigation of intercepts (and loadings, see Barendse et al., 2010, 2012; Woods & Grimm, 2011) invariance and also for an investigation of differential effects of gender within each of the treatment groups. This approach can easily be extended to the investigation of the effects of the treatment itself (i.e. with treatment group as the MIMIC variable). However, a main limitation of this approach is linked to the fact that the MIMIC approach assumes the invariance of the items' uniquenesses across levels of the MIMIC variables, an assumption that we know is violated in this application in relation to treatment group.

To estimate this MIMIC model, we started from the multiple group (defined on the basis of treatment groups) longitudinal model corresponding to the assumptions of partial strict measurement invariance (invariance of the configural model, factor loadings, items intercepts, and partial invariance of the items' uniquenesses, Model LT7p in Table 3). In this model, gender was included as a predictor of the latent means and items (according to the usual MIMIC sequence, with paths from gender to the factors and items freely estimated across time waves and groups. The results from these models are reported at the end of Table 3. For our simulated data, 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 at both time points (X5 and Y5). Allowing for direct effects of the predictor on item 5 in both the experimental and control groups, 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) = 5.922 (12), $p > .05$; and none of the ΔTLI, ΔCFI and ΔRMSEA show a decrement in fit]. We further investigated whether the direct effects of genders could be considered to be equivalent across treatment group in the context of an additional model and found this to be the case [$\Delta\chi^2$ (df) = 1.632 (2), $p > .05$; and the ΔTLI, ΔCFI and ΔRMSEA are all = 0]. Finally, we tested whether the effects of genders on the latent factor means could be considered to be equivalent across treatment group and found this not to be the case [$\Delta\chi^2$ (df) = 60 (4), $p < .01$; and the ΔTLI, ΔCFI and ΔRMSEA are all > .25]. Detailed results from the previous model assuming equivalent DIF across the experimental and control group but different effects of gender on the latent means (model Long-MIMIC-DIF-inv.inp) reveal that: (a) gender had no effect on neither the Anxiety nor Depression factors at *Pre-Test* ($\hat{\beta}$ = -0.029 to -0.156, *ns*); (b) *Post-Test* levels of Anxiety and Depression where did not differ according to gender in the control group($\hat{\beta}$ = 0.068 to 0.013, *ns*); (c) *Post-Test* levels of Anxiety tended to be substantially lower among females ($\hat{\beta}$ = -0.608, $p < .01$); (d) *Post-Test* levels of Depression did not significantly differ according to gender ($\hat{\beta}$ = -0.179, *ns*). Similarly, when the latent means (or rather latent intercepts of the regression of the latent factors on gender) are contrasted across the experimental and control groups, the results show that, when the latent means/intercepts are fixed to 0 in the control group at *Pre-Test*: (a) *Post-Test* levels of Anxiety do not significantly differ from Pre-Test levels in the Control group (-.159, *ns*); (b) *Post-Test* levels of Depression show a significant decrease from Pre-Test levels in the control group (-.403, $p < .01$), a decrease that is even more pronounced for the experimental group (-.692, $p < .01$); (c) *Post-Test* levels of Anxiety controlled for the effects of gender show a significant increase from *Pre-Test* levels in the experimental group (.690, $p < .01$). However, this last effect is really meaningless in itself as

it reflect the previously described treatment by gender interaction and should in fact be simply consider to reflect the intercept of the regression of Post-Test Anxiety on gender based on the following equation: *Post-Test* Anxiety = Intercept + b * gender = 0.690 + -1.673 * Gender (coded 0 for males and 1 for females). According to this equation, the Pre-Test to Post-Test latent mean differences observed in the experimental group reveal an increase in Anxiety levels of .690 for the males participants versus a decrease of .688 for females participants, fully consistent with the population model results and with the results from the more complex gender X treatment X time wave models considered previously.