Multigroup Confirmatory Factor Analysis as an alternative to (M)ANOVA

E. Maassen, M. A. L. M. van Assen, M. B. Nuijten, J. M. Wicherts

Traditionally, researchers who are interested in testing hypotheses on mean differences between groups employ (multivariate) analysis of variance ((M)ANOVA) methods. Univariate ANOVA is appropriate when analyzing means on a single outcome variable, whereas MANOVA is applicable when multiple outcome variables need to be compared simultaneously. MANOVA incorporates the correlations between the outcome variables (where ANOVA assumes these are zero), can identify which variables contribute most to group separation, and is able to look at the *set* of measures as they represent some underlying construct (Bray & Maxwell, 1985, p. 11). The use of (M)ANOVA methods appears to be on the decline within some fields in psychology (Troncoso Skidmore & Thompson, 2010), but researchers seem to remain largely unaware of more appropriate, informative, and statistically more powerful methods to analyze differences between groups (Breitsohl, 2019).

Although (M)ANOVA may be a convenient and suitable statistical method to identify group mean differences, it does not inform on the cause of these differences. That is, group means may differ because the construct is measured similarly across groups and these differences actually reflect different (mean) scores on the construct, or the group means may (at least partially) vary because the construct is measured differently across groups. Put differently and in statistical terms, (M)ANOVA cannot distinguish true mean differences on the construct from violations of *measurement invariance*. Measurement invariance relates to whether a set of variables has the same measurement properties in different conditions (e.g., over groups or over time; Meredith (1993)). Researchers who analyze constructs and wish to distinguish true group mean differences from violations of measurement invariance should therefore be careful when interpreting outcomes of (M)ANOVA (e.g., Camilli & Shepard (1987); Wicherts, Dolan, & Hessen (2005)).

The purpose of this article is twofold: first, we aim to explain and illustrate an

alternative method for researchers who are interested in comparing group differences on multiple measures of a (psychological) construct, i.e., multigroup confirmatory factor analysis (MGCFA). Second, we clarify what occurs when constructs are measurement non-invariant, but are analyzed with (M)ANOVA methods, which cannot account for this difference in measurement. [toevoegen: simulation study sentence + main message]

(Multivariate) Analysis of Variance: (M)ANOVA

Whenever a research question entails comparing mean scores between groups, (M)ANOVA methods can be used. In ANOVA, an F-test is performed that tests the null hypothesis that the population mean of the outcome variable is equal across all groups. In MANOVA, a generalization of the univariate ANOVA, a linear combination of multiple outcome variables is made that maximizes the differences between the groups (between-group variance) relative to the within-group variance. An omnibus test produces various test statistics (e.g., Wilks' lambda, Pillai-Bartlett trace) that test the null hypothesis that the population means of all variables for all groups are equal to one another. If this hypothesis is rejected, various follow-up steps exist. One such step is performing multiple univariate ANOVAs to investigate how much the groups differ on each outcome variable separately (which leads to a multiple comparisons problem and inflates the type I error rate). Alternatively, multiple outcome variables can also be summed together and subsequently analyzed as a univariate ANOVA (hereafter: *sum ANOVA*). This method does not inflate the type I error rate, but assumes all outcome variables are measured similarly and contribute equally to the overall construct. Another possibility is carrying out stepdown analysis, where one can use ANOVA to detect how much a given outcome variable adds to the between-group differences, starting with the outcome variable that was deemed most pertinent in discriminating the groups, followed by the second one while including the first outcome variable as a covariate, and so on (Stevens, 2009). A multivariate follow-up option to the omnibus test is discriminant function analysis, which classifies subjects into groups based on their scores on the outcome variables. Thus, MANOVA is able to investigate the relationships among variables and to select variables that contribute most to group separation or underlying theory. Because MANOVA examines multiple outcome variables together at once, it is a more powerful test to detect a difference in population means than performing separate ANOVAs with adjustment for multiple comparisons. However, ANOVA

methods are preferred if outcome variables are uncorrelated, or if outcome variables are negatively correlated within the groups being compared (Bray & Maxwell, 1985).

MultiGroup Confirmatory Factor Analysis (MGCFA)

MGCFA is a technique used to examine if and how groups differ in their measurement of a construct (or latent variable), and is able to estimate a latent mean difference between groups. MGCFA is an extension of Confirmatory Factor Analysis (CFA), a method used to analyze an a priori specified measurement model that relates measured outcome variables to one or multiple latent variables through regression (Jöreskog, 1971; Sörbom, 1974). MGCFA performs CFA in multiple groups to verify whether the measurement model has the same measurement properties in different groups, whereas (M)ANOVA assumes measures are comparable across groups. Figure 1 depicts an illustration of a prosocial behavior construct which underlies multiple observed outcome variables: other-, peer-, and self-reports. We assume that the covariation between these outcome variables occurs because of a common cause: prosocial behavior. In the figure, two measurement models are constructed: one for the control group on the left, and one for the experimental group on the right. Measurement errors ('e') for the outcome variables are also modelled. All estimated model parameters are depicted by Greek letters. The curved arrows that exit and reenter the same latent (ϕ) and error variables (ϵ) represent their respective variance estimates. The direct effects on the outcomes from the construct (λ) are called *factor loadings* and can (in a single-factor model) be interpreted as the correlation between the construct and outcome variable. Because we are interested in estimating mean differences between groups, we also include intercepts to the existing structure for both the latent and outcome variables, displayed by the two triangles (Kline, 2011). The intercepts of the outcome variables are depicted by ν. The κ parameters refer to the latent mean estimates (one for each group). These latter estimates are specifically relevant, since they provide the mean difference between groups that we are interested in. Performing MGCFA on the model in Figure 1 provides insight in how prosocial behavior is measured across the groups, and whether group differences reflect actual differences in prosocial behavior or differences in measurement. This makes MGCFA a more advantageous method to use than (M)ANOVA for researchers interested in testing mean differences between groups.

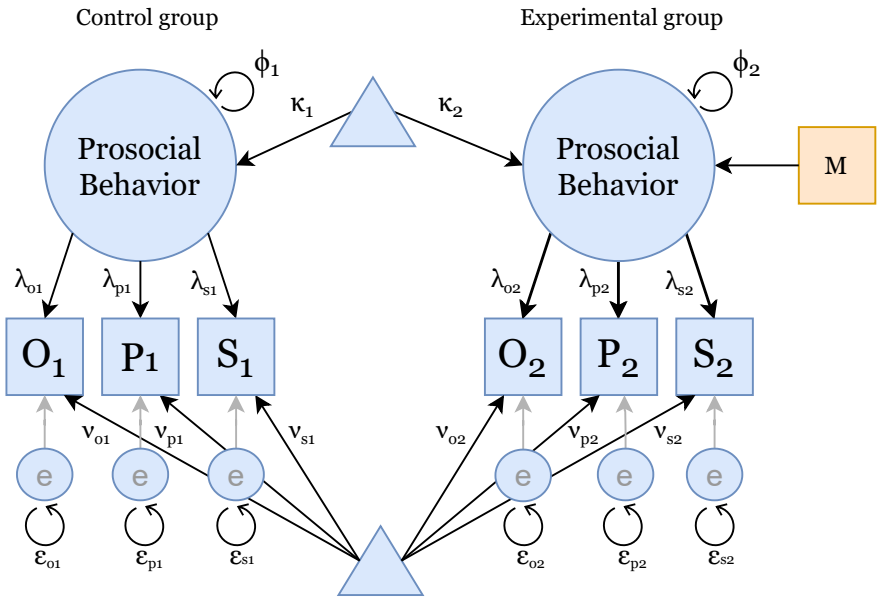

Figure 1: Measurement models for control group (l) and experimental group (r) with prosocial behavior as a factor and three measures of prosocial behavior: an other-report (O), peer-report (P), and self-report (S).

Measurement invariance

To ensure a latent mean differences between groups reflect actual differences on the construct, researchers must first establish *measurement invariance* between groups. Measurement invariance is tested stepwise by constraining sets of parameters in the measurement model to be equal across groups and comparing the fit of the model to the fit of the model in the previous step. Typically, measurement invariance is tested in four different phases. First, *configural invariance* tests whether the number of factors and the associated factor-indicator relationships are the same across groups. In our example, that means testing whether the three outcome variables load similarly on the prosocial construct in both groups. If configural invariance holds (i.e., the fit of the configural model is acceptable), the second phase tests *weak invariance*. We estimate a model where the factor loadings are set equal across groups (i.e., $\lambda_{o1} = \lambda_{o2}$, $\lambda_{p1} = \lambda_{p2}$, and $\lambda_{s1} = \lambda_{s2}$), and compare the fit of this model to the fit of the configural invariance model. If there is no statistically significant difference between the two models (i.e., the weak invariance model does not fit worse than the configural model), weak invariance holds. This means that the regression slopes of the three outcome variables are

(sufficiently) similar across groups. The third phase is testing for *strong invariance*, where we constrain the outcome intercepts to be equal across groups (i.e., $\nu_{o1} = \nu_{o2}$, $\nu_{p1} = \nu_{p2}$, and $\nu_{s1} = \nu_{s2}$), as well as still constraining the factor loadings. This model fit is compared to the fit of the weak invariance model, and if it fits equally well, strong invariance holds and the intercepts of the three outcome variables can be considered to be the same across groups. If factor loadings and intercepts are invariant over groups, i.e., if strong invariance holds, we can state that group differences are due to actual differences in prosocial behavior, and we can interpret the latent between-group mean difference. The fourth phase (*strict invariance*) constrains the residual variances of the outcomes across groups on top of the loadings and intercepts (i.e., $\epsilon_{o1} = \epsilon_{o2}$, $\epsilon_{p1} = \epsilon_{p2}$, and $\epsilon_{s1} = \epsilon_{s2}$). If strict invariance holds, the three outcomes are measured equally precise across groups. Note that generally, strict invariance is not necessary for estimating latent between-group differences; strong invariance between groups is sufficient.

Figure 2 depicts an example of strict measurement invariance with fictitious estimates. All loadings and intercepts are equal in the control and experimental group, and the estimated latent between-group difference is $\kappa_2 = 0.2$ (which in this example can be interpreted as standardized mean difference Cohen's d). The observed group differences ($\mu = \nu + \lambda \times \kappa$) are dissimilar for each outcome variable: the difference in the self-report (0.12) is larger than that of the peer-report (0.10), which is larger than that of the other-report (0.08). Larger loadings indicate that the observed variable is more indicative to the construct. Moreover, as we have a (standardized) single-factor model, a loading λ is also a correlation, and its square λ^2 is the amount of variance in the outcome variable that is explained by the latent construct of prosocial behavior, with $1 - \lambda^2$ being equal to the unexplained or error variance of the outcome variable. Because there is (strict) measurement invariance, we can interpret the group differences as actual differences in prosocial behavior.

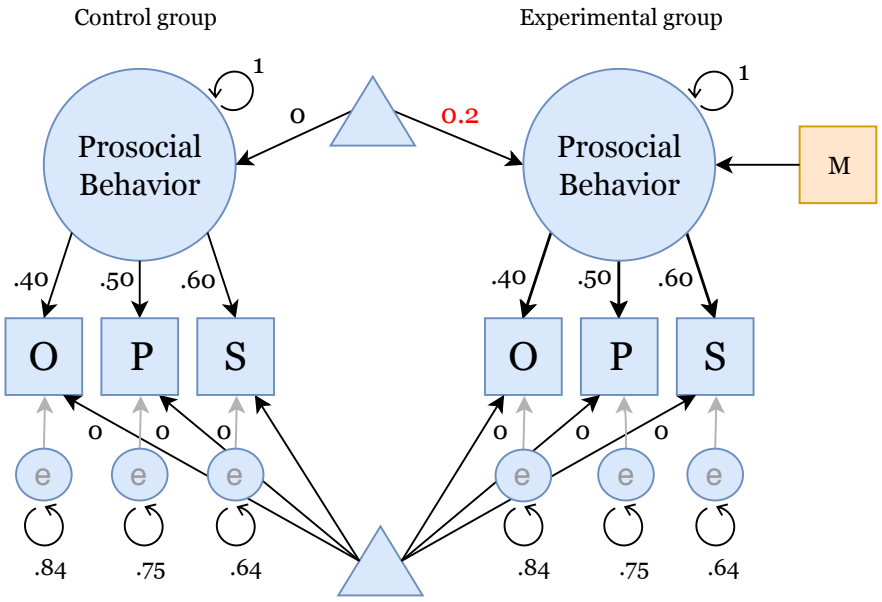

Figure 2: Measurement models for control group (l) and experimental group (r) with prosocial behavior as a factor and three measures of prosocial behavior: an other-report (O), peer-report (P), and self-report (S). The loadings, intercepts, and residual variances are invariant across groups, implying strict measurement invariance.

Measurement non-invariance

MGCFA can also give insight on violations of measurement invariance, and thereby explain observed mean differences in outcomes variables that are not the result of (true or) latent mean differences in prosocial behavior. It could occur that a manipulation in the experimental group inadvertently invokes attention, which in turn invokes social desirable behavior on the self-report measure in the experimental group (S_2). This effect is depicted in Figure 3. In this case, the observed mean differences between the groups on the self-report variable cannot fully be explained by group differences in prosocial behavior. Specifically, these differences can also be caused by differences in the loading or intercept of the self-report variable; we call this *measurement non-invariance*.

In Figure 3, the intercept of the self-report measure in the experimental group has increased from 0 to 0.3 due to the effect of an additional latent variable, resulting in a mean difference on the self-report measure equal to $\mu = 0.3$ (because of measurement non-invariance) $+ 0.5 \times 0.6$ (effect of manipulation on the latent construct) $= 0.6$. Since (M)ANOVA methods

only test group mean differences on the observed outcome variables and cannot disentangle differences due to true mean differences or measurement non-invariance, (M)ANOVA would estimate the group mean difference on the self-report measure as 0.6, which is 0.3 higher than the difference on the underlying construct. Another possibility is that the estimated group mean difference on the outcome variable is 0 (if the intercept is −0.3, rather than 0.3), in which case the true group mean difference (0.3) is completely erased by measurement non-invariance. Contrary to (M)ANOVA, MGCFA disentangles and estimates both (the effect due to latent difference as well as the effect due to measurement invariance), thereby providing a correct interpretation of potential observed group differences in outcome variables. [Note: omdat er hier measurement non-invariance is, klopt de latent mean estimate (misschien) ook niet.]

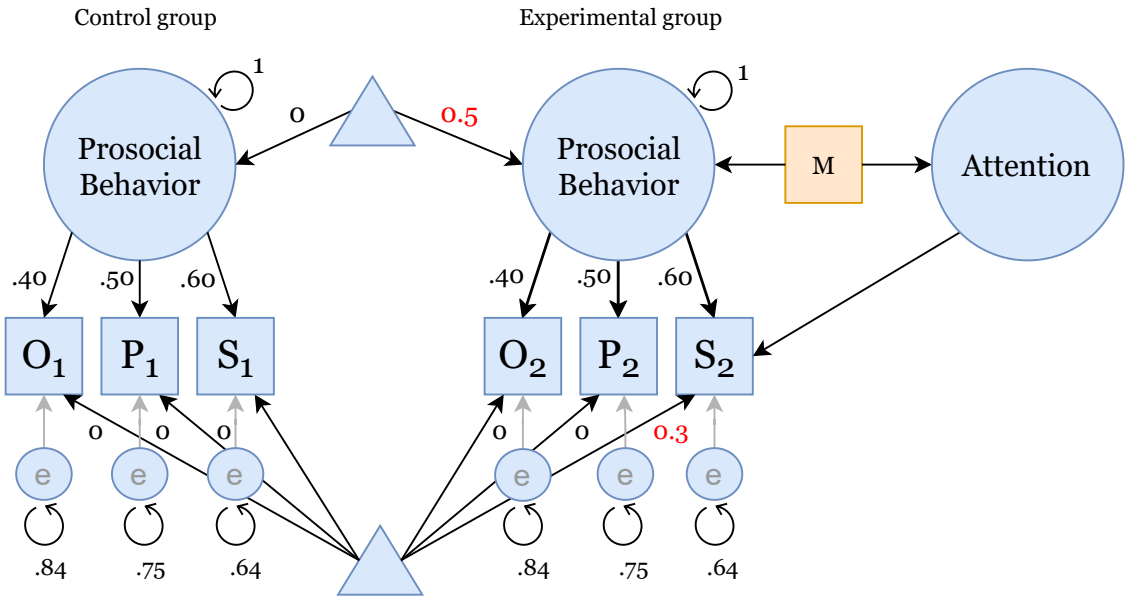

Figure 3: Measurement models for control group (l) and experimental group (r) with prosocial behavior as a factor and three measures of prosocial behavior: an other-report (O), peer-report (P), and self-report (S). The intercepts between groups are not the same, implying measurement non-invariance.

Previous literature

Neither MGCFA nor (M)ANOVA are novel methods, and much research has been conducted on their performance under certain conditions. MGCFA is more powerful in

detecting mean differences than MANOVA in situations of invariance (van Smeden & Hessen (2013)) and factor loading non-invariance, especially when effect sizes are small ($\kappa \leq 0.5$, Hancock, Lawrence, & Nevitt (2000)). Especially non-invariant intercepts can have a large impact on the latent mean estimation, which reduces the probability of drawing correct statistical conclusions based on a latent mean difference test (De Beuckelaer & Swinnen, 2011). This finding is problematic but not surprising, since intercept non-invariance will bias the latent mean difference for an entire group (i.e., across all scores on the latent construct), whereas bias due to loading non-invariance might only affect part of the group (i.e., is dependent on someone's score on the latent construct). We may also reasonably assume an effect differs for an entire group, instead of only for people with a certain latent variable score. There might be conditions under (partial) intercept non-invariance in which MGCFA is able to provide an accurate latent mean estimate where (M)ANOVA fails to do so. An additional issue is that many of these studies, as well as those focused on measurement invariance, focus on sample sizes of at least 200 per group. This is an unrealistically large estimate for psychological research, where the median sample size related to t, F, and r statistics is $N = 62$ (Hartgerink, Wicherts, & Van Assen (2017)). For studies that do investigate power to detect measurement invariance with smaller sample sizes, either no comparison is made to the estimation of the latent mean difference or (M)ANOVA methods (e.g., Meade & Bauer (2007)). Our aim is to show that MGCFA is a suitable method for many research designs in which group mean differences are of interest, as well as show that (M)ANOVA is not a suitable method when there is measurement non-invariance, and that MGCFA can provide a good alternative.

This study

Our study serves multiple purposes. Firstly, we investigated under which conditions it is realistic to estimate a MGCFA model. With MGCFA, the loglikelihood of the data given the specified model is maximized, but for some model and data combinations non-convergence can arise. We were interested to find out the minimal settings (e.g., smallest sample size) under which estimating a MGCFA model is still feasable. To this end, we checked convergence rates of the model under all conditions and eliminate conditions under which estimating a MGCFA model is attainable, before moving on to the next step. Secondly, we investigated how well MGCFA and (M)ANOVA perform under measurement invariance. We compared

all methods in terms of type I error rate, the statistical power to find a latent or observed difference between groups, and effect size estimation. This can help give researchers an idea on which method to use under ideal conditions, when latent constructs are comparable across groups. Finally, in Study 2 we checked how both MGCFA and (M)ANOVA performed when there is intercept non-invariance between groups. We checked (M)ANOVA's performance in terms of effect size estimation when there is measurement bias in the construct, as well as type I error rate and statistical power to find measurement non-invariance for MGCFA. Additionally, we checked which fit measures could accurately recover the correct partial invariance model under intercept non-invariance.

Method Study 1

In the first simulation study we compared the performance of multiple ANOVA methods to MGCFA under measurement invariance between groups, in terms of statistical power to detect a mean difference and effect size estimation. The MGCFA model will be used for the generation of data [hier is meer verantwoording voor nodig denk ik?]. The data were generated by using the *MASS* (version 7.3-51.4, Venables, Ripley, & Venables (2002)) and *psych* package (version 1.8.12, Revelle (2020)) in *R* (version 3.6.1, R Core Team (2019)). More information on the methods we used in the simulation studies and the simulation code can be found in Appendix A (OSFlink) and at OSFlink respectively.

Data generation

Data for two groups on three continuously measured observed variables were generated, assuming one underlying latent variable in the MGCFA approach, as depicted in Figure 2. For an individual j in group g, their observed score on the three variables are modelled according to the following measurement equation:

$$y_{gj} = \nu_g + \lambda_g \, \eta_{gj} + \epsilon_{gj},$$

where ν is a vector of three intercepts for group g, λ is a vector of three factor loadings for group g, η is the subject-specific latent variable score, and ϵ is the subject-specific residual. Subject-specific latent variable scores (η) and residual variances for three observed variables (ϵ) were drawn from the normal distribution: $\eta \sim \mathcal{N}(\kappa_g, \phi_g)$ and $\epsilon_{1gj} \sim \mathcal{MVN}(0, \Theta_g)$, where κ and ϕ represent the mean and variance of the latent variable respectively. Θ_g is a diagonal matrix containing the residual variances for the observed variables ϵ, which were generated

equal to $1 - \lambda^2$.

Since we were interested in the performance of the models under measurement invariance, the loadings and intercepts in all conditions were generated to be equal across the groups (i.e., $\lambda_{g_1} = \lambda_{g_2}$ and $\nu_{g_1} = \nu_{g_2}$). All intercepts ν were chosen to be equal to 0. Additionally, within all conditions we kept the sample sizes of the groups the same ($n_{g_1} = n_{g_2}$). We set the mean of the latent variable in the first group to zero so it serves as the reference group, and generated latent mean differences for group 2.

Factors we varied were the sample size per group (25 to 500 per group), the magnitude of the latent mean difference (Cohen's d 0 to 1.0), and loadings, for a total of $10 \times 6 \times 27 = 1620$ conditions (see Appendix A for an overview: OSFlink). The total number of iterations for each condition was set to 1000.

Analyses

MGCFA. To evaluate the MGCFA method, we fitted two separate single-factor models to the generated covariance matrix (3×3) and mean vector (3×1) [meer uitleg?] in each iteration: one where the latent mean in group 2 was free to be estimated (i.e., the correct model), and one where the latent mean in group 2 was fixed to zero (i.e., the incorrect model). In both models we constrained the loadings and observed intercepts to be equal across groups (i.e., strong invariance). We compared the fit of these two models through a likelihood-ratio test with one degree of freedom (i.e., κ_{g_2}), tallying the number of times the latent mean in group 2 showed a statistically significant difference ($\alpha = .05$) from the mean of zero in group 1 [noncentrality parameter uitleggen?]. We used the maximum likelihood estimator in the lavaan package (version 0.6-5, Rosseel (2012)).

Because we used a standardized solution, the latent mean estimate of group 2 (κ_{g_2}) can be interpreted as the standardized mean difference between groups (Cohen's d). To investigate coverage of the effect size, we tallied how often the expected latent mean difference between groups (κ) fell within the 95% confidence interval of the calculated mean difference.

(M)ANOVA. Whereas evaluating the MGCFA methods revolves around group differences in latent means, for the (M)ANOVA methods we are interested in the group differences in observed means. For all methods, we used the F-test to tally the number of times the null hypothesis of no mean differences between groups was rejected. To evaluate the ANOVA

method we used a one-way F-test for each of the three outcome variable separately to examine statistically significant differences between group means. To evaluate the sum ANOVA method, we calculated a sum score of the three observed variables for all participants in both groups and performed a one-way F-test on this sum score. Note that in these two cases, the F-test reduces to the t-test because only one dependent variable is analyzed at a time. To evaluate the MANOVA, we used the F-test to simultaneously compare the means of the three dependent variables across groups.

For both the separate ANOVAs and the sum ANOVA we calculated the standardized mean difference between groups, i.e. Cohen's d, through the following formula:

$$d = \frac{\mu_2 - \mu_1}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{n_1 + n_2 - 2}}$$

To investigate coverage of the effect size, we tallied how often the expected observed mean difference between groups (μ) fell within the 95% confidence interval of the calculated mean difference.

Results Simulation Study 1

Minimal conditions for MGCFA

One of our aims of the first study was to investigate under which minimal conditions it is feasible to use the MGCFA method to estimate a latent mean difference. To this end, we filtered out conditions that did not meet the criteria related to convergence and coverage. Some of the 1620 conditions did not achieve acceptable model convergence rates; in 91 conditions, the average percentage of nonconvergence over all 1000 iterations was lower than 80%. Some conditions (MGCFA 853, Sum ANOVA 201) did not have acceptable coverage estimates of between 94% and 96% of the latent or observed mean estimate. Other conditions (MGCFA 96, Sum ANOVA 739) did have acceptable coverage rates, but such large confidence intervals that the average latent or observed mean estimate deviated more than 0.1 from the population mean estimate. Note that multiple of these issues could and did often occur within one condition. We filtered out all conditions that did not meet the criteria for convergence and coverage. The following results are based on the remaining 580 conditions.

Results show that it is almost always infeasible to use a MGCFA model with a sample

size of 25 per group; the model only has accurate convergence and latent mean accuracy when factor loadings are high (all ≥ 0.7, or an eigenvalue of 1.47) regardless of latent mean difference, and in these scenarios latent mean estimates are slightly positively biased. Note that extending the scale with more items increases the amount of variance in the latent variable that is explained by the observed variables, and thus the eigenvalue. In doing so, the individual factor loadings can be lower than the 0.7 treshold while still upholding the same amount of explained variance. Moreover, it is possible to estimate a MGCFA model when correlations between variables are small (i.e., $r = 0.20$), but the minimum sample size needed is then 200 per group. Bare in mind that these correlations assume perfect reliability of the observed scores because they are free of measurement error. In cases where observed scores are less reliable, observed correlations need to be larger. In short, the MGCFA model is quite accomodating to small sample sizes ($N = 50$ per group) and small correlations between variables ($r = 0.20$), but at least one of these needs to be relatively high to get accurate estimations in MGCFA.

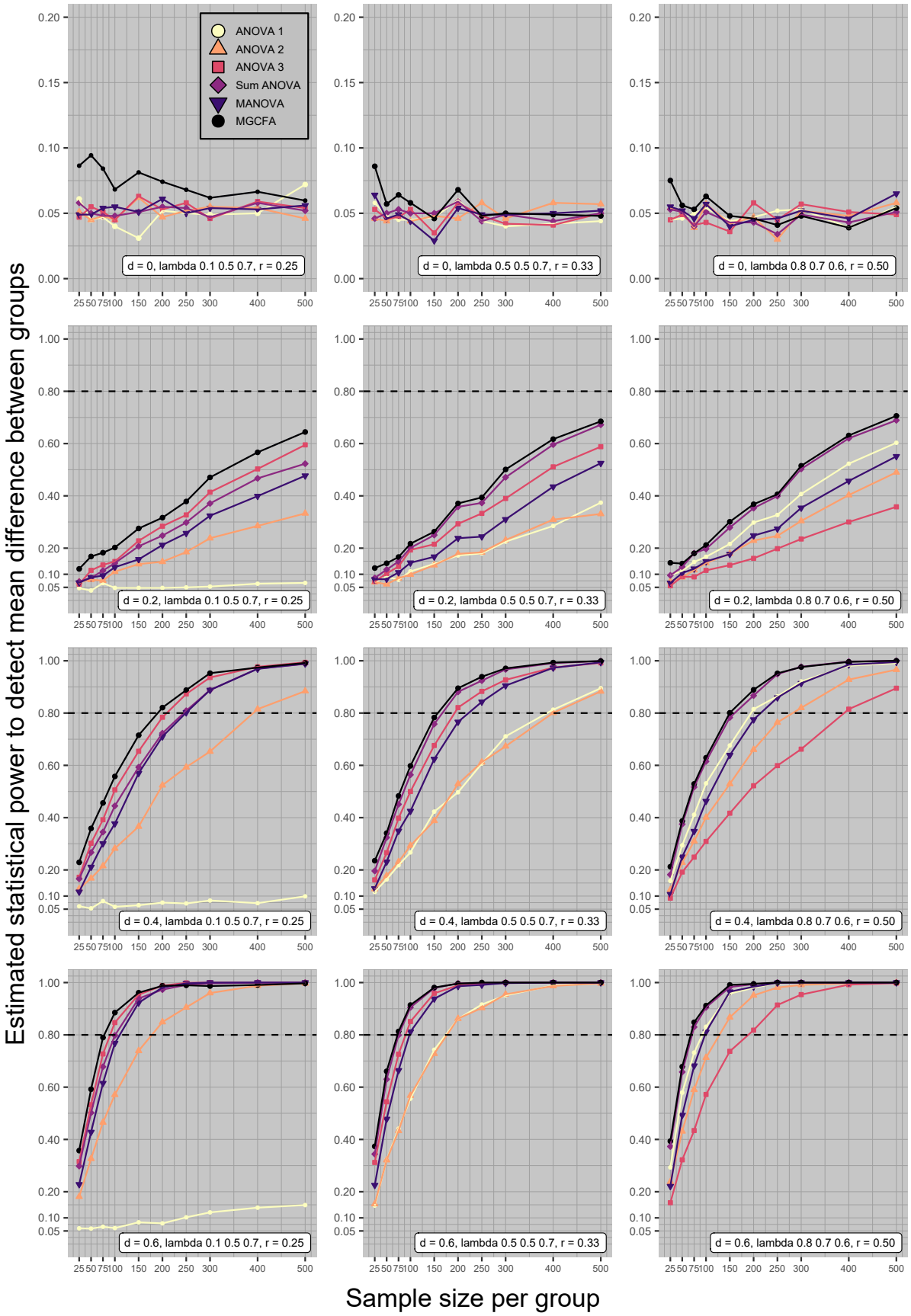

Figure 4: Statistical power to find a mean difference of the MGCFA and ANOVA methods under differing effect sizes (d) and an average correlation between observed variables of r = 0.37. The dashed line indicates statistical power of 0.80.

Type I error

We investigated the type I error in conditions where there was no mean difference between groups, for all methods; the results are depicted in the top left plot of Figure 4 for conditions with an average correlation among variables of $r = 0.37$. Type I error rate for all method remained within Bradley's liberal robustness criterion (i.e., $\alpha \pm 0.5\alpha$, so ≥ 0.25 and ≤ 0.75, Bradley (1978)), except for MGCFA in three conditions. These were conditions in which two out of three factor loadings were very small (i.e., 0.1 and 0.3), resulting in type I error rates of 0.076 and 0.077. For all methods, type I error overestimation was negatively associated with eigenvalue (MGCFA r = - 0.29, Sum ANOVA r = - 0.64, MANOVA r = -0.69) and, to a lesser extent, sample size (MGCFA r = - 0.16, Sum ANOVA r = - 0.33, MANOVA r = -0.32), meaning that the more the variance the factor explained and the higher the sample size, the more Type I error rate tends to the nominal rate of .05.

Statistical power

Statistical power to detect a latent or observed mean difference is a function of multiple properties; correlations between observed variables, effect size (i.e., group mean differences), and sample size. It is possible to trade-off these properties: as the correlation between observed variables increases, the required effect size and sample size can decrease without a loss in power. Figure 4 displays the statistical power to find a group mean difference of the methods in a situation where the average correlation among variables is $r = 0.37$, for multiple effect sizes.

In almost all conditions, the statistical power of the MGCFA LR test exceeded the power of all (M)ANOVA F tests. In conditions where this was not the case, the F test power estimate never exceeded that of the LR test with more than 0.02. As is shown in Figure 4, MGCFA clearly outperforms all (M)ANOVA methods except the sum score ANOVA method, which is the case because there is measurement invariance in all conditions [meer uitleg nodig]. This result shows that in cases of measurement invariance at the loading and intercept level, power to estimate a mean difference between groups is approximately equal between the LR test and the F-test of the sum score. However, where ANOVA assumes measurement invariance, MGCFA can actually test for it, which is the focus of the next study.

As is shown in Figure 4, statistical power never exceeds more than 0.80 in any of the

methods if the effect size is small ($d = 0.2$) when the average correlation between the three observed variables is $r = 0.37$. [Hier nog over schrijven?]

Conclusion study 1

Under measurement invariance, it is feasible to estimate a MGCFA model if either the correlations between observed variables *or* sample sizes per group are relatively large, with an absolute minimum of $r = 0.20$ and $N = 200$ per group, or $r = 0.40$ and $N = 50$ per group. However, we note that statistical power in these circumstances, as well as when the effect size is small, will not reach the standard treshold of .80. The type I error rate was nominal in all (remaining) conditions and for all methods. Statistical power of the LR test outperformed power of the MANOVA F-test in all conditions, and was relatively equal to the power of the sum score ANOVA F-test. When there is measurement invariance, sum score ANOVA and MGCFA perform equally well. However, ANOVA methodes always assume there is measurement invariance, whereas MGCFA provides more information and can test for non-invariance, which is the focus of the next study.

Study 2

In the second simulation study our aim was to compare the performance of (M)ANOVA methods to MGCFA, but under measurement non-invariance. Researchers faced with the option to do a measurement invariance check can do so in different ways. A common practice is to check for configural invariance first, then loading invariance, and as a final step (if estimating a latent mean difference between groups is the goal), intercept invariance. If intercept invariance does not hold, multiple options are available. One of these options is checking the modification indices from the model that restricts all intercepts to be equal (i.e., the incorrect model). The modification indices test the constrained parameters seperately in univariate score tests. A statistically significant modification index for a certain parameter indicates that that parameter may have to be freed to ensure better model fit. If intercept invariance is rejected, but the researcher does not know in which intercept(s) bias exists, another possibility is to fit many different models to the data, and check the absolute and relative fit indices to see which of these models indicates the best fit. An issue with this approach is that there are often many different models to fit; in a situation with one factor and three indicators (in Table 1), it would be possible to estimate 14 different models that either check for full (model 1 and 2) or partial (models 3 to 14) intercept invariance. Note

Table 1: Possible (partial) strong invariance models to estimate in a one-factor three-indicator model

No.	Model	Factor intercepts	Factor mean group 2
1	Strong invariance	y1, y2, y3 invariant	Fixed at 0
2	Strong invariance	y1, y2, y3 invariant	Free to estimate
3	Partial strong invariance	y1, y2 invariant	Fixed at 0
4	Partial strong invariance	y1, y2 invariant	Free to estimate
5	Partial strong invariance	y1, y3 invariant	Fixed at 0
6	Partial strong invariance	y1, y3 invariant	Free to estimate
7	Partial strong invariance	y2, y3 invariant	Fixed at 0
8	Partial strong invariance	y2, y3 invariant	Free to estimate
9	Partial strong invariance	y1 invariant	Fixed at 0
10	Partial strong invariance	y1 invariant	Free to estimate
11	Partial strong invariance	y2 invariant	Fixed at 0
12	Partial strong invariance	y2 invariant	Free to estimate
13	Partial strong invariance	y3 invariant	Fixed at 0
14	Partial strong invariance	y3 invariant	Free to estimate

that when a (latent) mean difference between groups is expected, a researcher would fit the even numbered models, whereas the uneven numbered models would be fitted if there is no expected (latent) mean difference between groups. With more indicators, the possible models to fit rise substantially. Our aim with the second study is to investigate which choices researchers who would like to estimate a latent mean difference between groups but are faced with intercept non-invariance in their MGCFA model could make. We also checked the sensitivity of multiple fit measures, to find out which would be suitable to consult for model selection.

Additionally, we investigated MGCFA's power to detect intercept bias. Since (M)ANOVA does not allow for testing measurement invariance, we are unable to compare MGCFA's power to that of (M)ANOVA. However, we expect the (M)ANOVA estimated mean difference between groups to be strongly over- or underestimated in cases of intercept non-invariance, and will compare the latent mean differences from MGCFA to the observed mean differences from (M)ANOVA.

Method Study 2

In study 2, we fitted all models in Table 1 to the generated data and extracted relevant information. Multiple fit measures were extracted to find out if they correctly specified the correct model given the data. With this setup, there are four different possibilities in

simulating data, and four different correct models given each of these situations. When there is intercept invariance and no expected group difference, model 1 is the correct model. When there is intercept invariance and an expected group mean difference, model 2 is the correct model. Note that both these models were used in Study 1. When there is intercept non-invariance (which is always simulated in the third intercept) but no expected group difference, model 3 is the correct model. Here, the biased indicator is freed so it can be estimated differently in each group. When there is intercept non-invariance and an expected group difference, model 4 is the correct model. Our aim is to investigate which indices or fit measures indicate the correct model to choose in all four situations.

To find out whether the most restricted models (i.e., model 1 and 2) correctly indicated the third intercept to be biased, we saved the modification indices for model 1 and model 2 in situations of intercept bias. The modification indices perform univariate tests on restricted parameters to see if model fit improves when the restriction is freed. We tallied the number of times the modification indices correctly identified the third intercept as non-invariant across groups.

To investigate model fit, we estimated absolute and comparative fit indices for all models. Absolute fit indices compare the observed correlations to the model-implied correlations to assess model fit, whereas incremental indices compare the fit of the estimated model to a null model that generally does not fit the data well. The χ^2 test can be considered as the main measure to assess absolute model fit, but in measurement invariance testing it is also often used to test the relative fit of two models. However, the χ^2 test is sensitive to sample size; as sample size increases, the differences between the observed and model-implied data becomes smaller, increasing the probability of rejecting the null hypothesis of perfect fit. Therefore it is important to consider other fit measures as well. Other examples of inappropriate fit measures are the TLI and GFI, because they do not penalize for complexity or are dependent on the size of the correlations between the observed variables. We present the results of these inappropriate fit measures in appendix B, but will only discuss the appropriate fit measures next.

The absolute fit measures we included are the RMSEA, SRMR, AIC, and BIC. The RMSEA, or Root Mean Square Error of Approximation (Browne & Cudeck, 1992; Steiger, 2016), considers exact fit of the model to the data to be unrealistic, and instead estimates

approximate fit. The RMSEA uses the $\chi^2 - df$ value corresponding to an RMSEA of .05 (i.e., the noncentrality parameter)[weg?], sample size, and degrees of freedom to calculate approximate fit. The index ranges from 0 to 1, where values $< .05$ indicate close fit. The SRMR, or Standardized Root Mean Square Residual (L.-T. Hu & Bentler, 1995), is the standardized difference between the observed correlations from the data and the model-implied correlations, with a value of zero indicating perfect fit. Values < 0.8 are considered a good fit to the data (L. Hu & Bentler, 1999). The Akaike Information Criterion, or AIC (Akaike, 1974) is an absolute fit measure that estimates how well the model fits the data and penalizes the number of parameters. The likelihood of a model fit can be increased by adding parameters, which can lead to overfitting. AIC corrects for this by penalizing complexity, meaning less complex models are preferred. The values of the AIC in itself can not be interpreted, but it is possible to compare model fit by comparing the AIC estimates; the lower the AIC value, the better the model fit. The Bayesian Information Criterion, or BIC (Schwarz, 1978), is similar to the AIC in that it also assesses model fit based on the likelihood of the data and model complexity. However, BIC penalizes more strongy than AIC for model complexity, meaning that models with more parameters will have a higher estimate and worse fit.

For relative fit measures we included the CFI, or Comparative Fit Index (Bentler, 1990). The CFI compares the fit of the model to the data to the fit of an independent or null model. The traditional independence null model used in the package we use (lavaan, Rosseel (2012)) is not nested in our (multiple group) models, which means the traditional CFI estimate in this study is inappropriate to use (Widaman & Thompson, 2003). As such, we estimated a null model with group invariant intercepts and residuals and mutually uncorrelated observed variables (see OSFlink, model 0A in Widaman & Thompson (2003)). We named this fit measure CFI_w. The CFI ranges from 0 to 1, with a value of 0.95 or higher indicating acceptable model fit (Schreiber, Nora, Stage, Barlow, & King, 2006). Results for the traditional CFI measure are displayed in Appendix B.

We estimated type I error and power to detect intercept bias for MGCFA by comparing the fit of model 1 to model 3 (when there is no group mean difference) and the fit of model 2 to model 4 (when there is a group mean difference). We estimated type I error and power to detect a latent mean difference across groups in the presence of intercept bias for MGCFA, by

comparing the fit of model 3 to model 4, where the only difference between models is the freed latent mean in group 2. We did not calculate statistical power to detect an observed mean difference for (M)ANOVA methods, since this power will be conflated with the simulated intercept bias, and will thus be over- or underestimated. We did estimate the observed mean difference between groups for the (M)ANOVA methods.

Data generation

Data generation was identical to that of Study 1 with the exception of the intercepts ν. Again, data for two groups on three continuously measured observed variables were generated, assuming one underlying latent variable in the MGCFA approach, as depicted in Figure 2. The intercepts were chosen equal to 0 for the first and second indicator in both groups ($\nu 1_{g_1} = \nu 1_{g_2}$ and $\nu 2_{g_1} = \nu 2_{g_2}$), and for the third indicator in the first group. Intercept differences were simulated in the third indicator for the second group, with a range of 0.2 to 1 ($\nu 3_{g_1} \neq \nu 3_{g_2}$). Additionally, within all conditions we kept the sample sizes of the groups the same ($n_{g_1} = n_{g_2}$), as well as all the factor loadings. We set the mean of the latent variable in the first group to zero so it serves as the reference group, and generated latent mean differences for group 2.

Factors we varied were the sample size per group (25 to 500 per group), the magnitude of the latent mean difference (Cohen's d 0 to 1.0), the magnitude of the factor loadings, the magnitude of the third indicator intercept in group 2 ($\nu 3_{g_2} = 0 to 1.0$), for a total of $10 \times 6 \times 12 \times 6 = 4320$ conditions (see Appendix A for an overview: OSFlink). The total number of iterations for each condition was set to 1000.

Analyses

In cases where we simulated intercept non-invariance between groups, we investigated power to detect intercept differences between groups by performing a 1 *df* LRT test to compare goodness-of-fit between model 1 and model 3. The 1 *df* difference in this case is the parameter of the third intercept in group 2, which we free in model 3. We also investigated goodness-of-fit between model 3 and model 5 with a 1 *df* LRT test, to detect which fit measures prefer a less restricted model (i.e., model 5) to a more restricted one (i.e., model 3). The 1 *df* difference is the second intercept in group 2, which is unbiased and which we free in model 5. We tallied the number of times the 1 *df* LRT test was statistically significant. Note

that the null hypothesis states that the more restricted model fits equally well, meaning that a statistically non-significant result is desirable in this case.

To assess power to detect a group mean difference for MGCFA we performed three separate 1 *df* LRT tests: in the case of no intercept bias we compared the fit of model 1 to model 2, where all intercepts are restricted to be equal. In the case of intercept bias we first compared model 3 to model 4, where the third intercept is freed in both models. Lastly, we compared model 5 to model 6, where the second and third intercept are freed in both models. The 1 *df* difference in these three cases is the parameter of the latent mean in group 2, which we free in model 2, 4, and 6. We again tallied the number of times the 1 *df* LRT test was statistically significant.

Since (M)ANOVA methods are unable to investigate measurement non-invariance, we did not check power to detect a group mean difference for these methods. However, since we expect the observed mean difference in these methods to be conflated with the intercept bias, we did plot the estimated latent mean difference (for MGCFA) to the observed mean difference (for ANOVA). For both the separate ANOVAs and the sum ANOVA we calculated the standardized mean difference between groups, i.e. Cohen's *d*, through the following formula:

$$d = \frac{\mu_2 - \mu_1}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{n_1 + n_2 - 2}}$$

To investigate coverage of the effect size, we tallied how often the expected observed mean difference between groups (μ) fell within the 95% confidence interval of the calculated mean difference.

Results Simulation Study 2

Modification indices and fit measures

Modification indices indicate whether parameters that were restricted to be equal across groups should be freed to vary across groups to improve model fit. We estimated the number of times the modification indices correctly identified the biased intercept as non-invariant in model 1 (in situations where there is no group mean difference) and model 2 (in situations where there is a group mean difference). [opnieuw runnen of nooit opgeslagen?]

For all conditions we considered which model was chosen to be the best fitting model for each fit measure, of which results are displayed in Figure X for various levels of group mean difference and intercept bias. Results show that the χ^2 test, the SRMR, and the TLI tend to favor the less parsimonious models (i.e., model 12, 13, and 14 in Table 1), meaning they are often unable to identify the correct model when other, less parsimonious models are presented. For the AIC, BIC, RMSEA, and CFI, the proportion of times the correct model is indicated as the best fitting model increases as sample size increases. This is also the case for intercept bias; as intercept bias becomes larger, the probability of selecting the correct model increases. A notable exception to this result is when the intercept bias is 0.2 (the second column of Figure X); in these situations, the more restricted models (i.e., model 1 or 2 in Table 1) often indicate a better fit, because the intercept bias is relatively small. Finally, the proportion of times a correct model is selected also deceases as the latent group differences increase.

Of all selected fit measures, the AIC, BIC, and RMSEA perform the best. In cases where the group mean difference or intercept bias are small, the BIC outperforms the other fit measures. As group differences and intercept bias increase, the AIC and RMSEA select the correct model most often. Since the proportion of correct model selection in most cases is lower than 75%, we recommend using a combination of the AIC, BIC, and RMSEA when examining goodness of fit of your models.

Type I error

We investigated the type I error for finding a latent or observed group mean difference for all methods; the results are displayed in Figure X for conditions with an average correlation among variables of $r = 0.37$. In cases without intercept bias (panel A), the type I error rate for all methods remained within Bradley's liberal robustness criterion (i.e., $\alpha \pm 0.05\alpha$, so ≥ 0.025 and ≤ 0.075, Bradley (1978)). In cases with intercept bias, the type I error rate for all (M)ANOVA methods increasingly exceeds Bradley's criterion as intercept bias and observed group mean differences increase. As in Study 1, type I error overestimation was negatively associated with eigenvalue and sample size for all methods.

For MGCFA, we examined the type I error for finding intercept bias, of which results are displayed in Figure X. Over all conditions and all levels of intercept bias, the type I error rate remained within Bradley's criterion.

Statistical power

Since statistical power to detect an observed mean difference for (M)ANOVA methods is conflated with the magnitude of intercept bias, we only display results for the statistical power of MGCFA to detect a latent mean difference in situations with intercept bias next. Statistical power depends on the correlation between variables, the magnitude of the group mean differences, and sample size.

Effect size

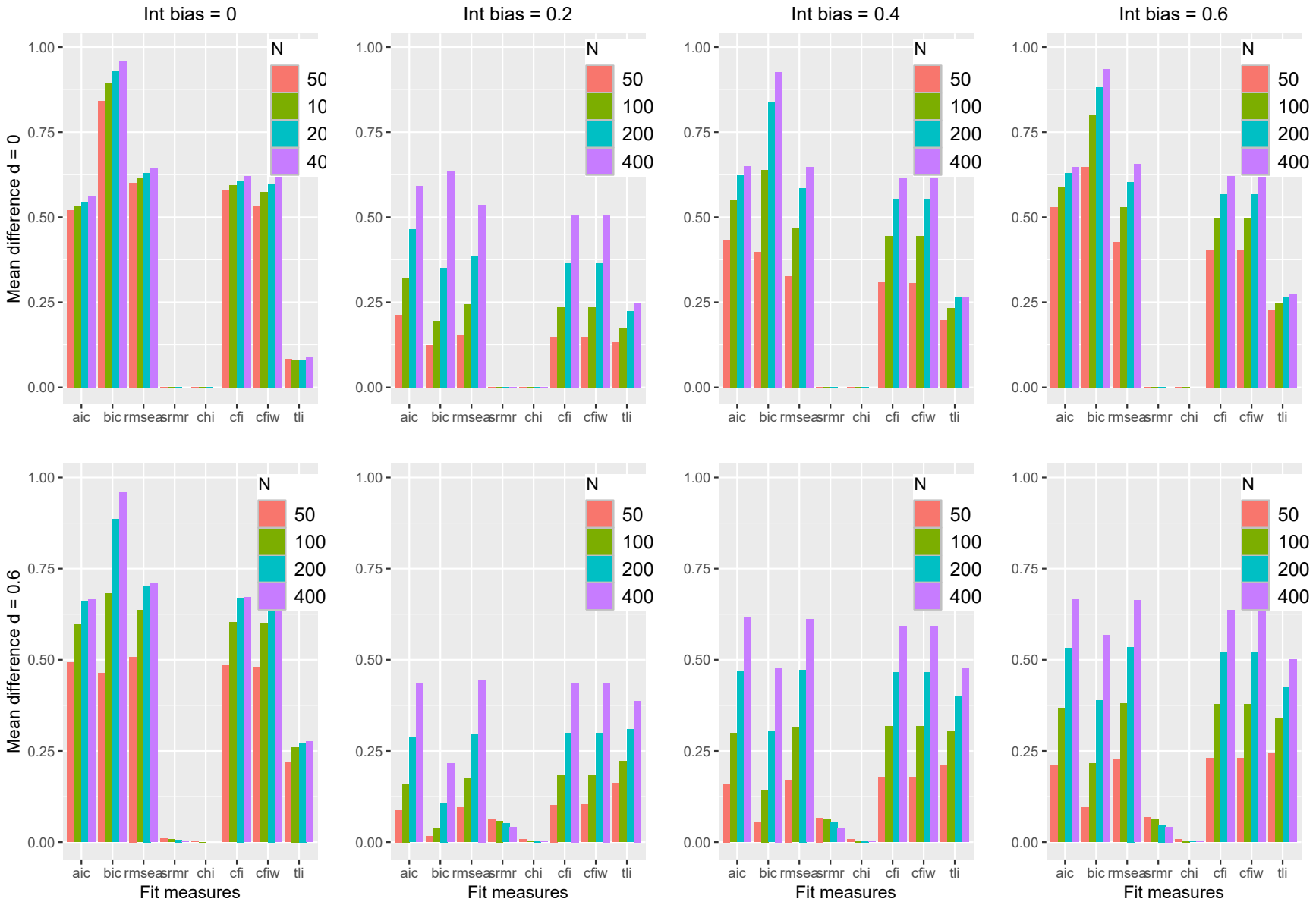

Figure 5: Proportion of times the correct model was chosen as the best fitting models for multiple fit measures.

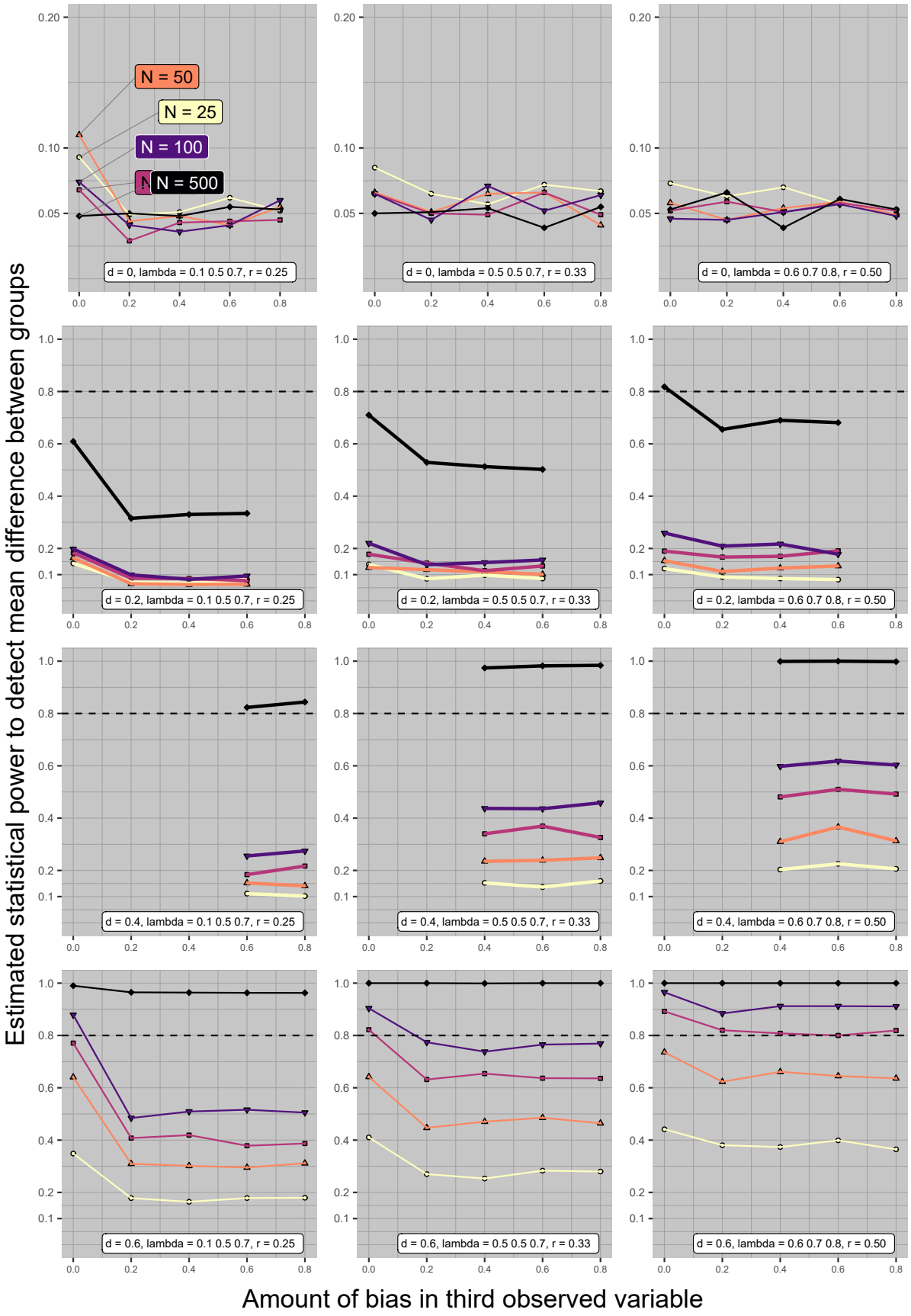

Figure 6: Statistical power to find a mean difference of the MGCFA methods under uniform bias, for differing effect sizes (d). The dashed line indicates statistical power of 0.80.

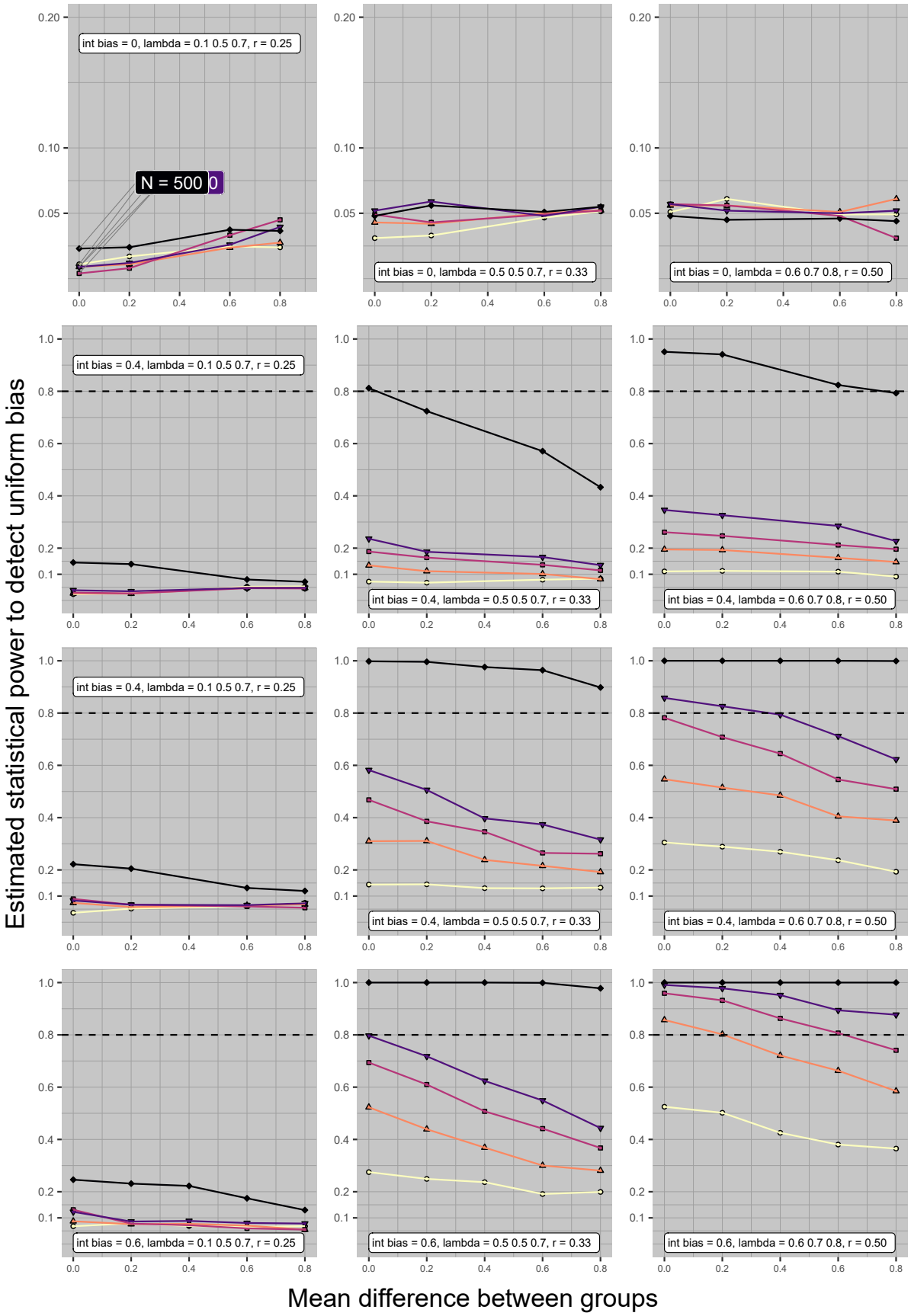

Figure 7: Statistical power to find uniform bias of the MGCFA methods for differing effect sizes (d). The dashed line indicates statistical power of 0.80.

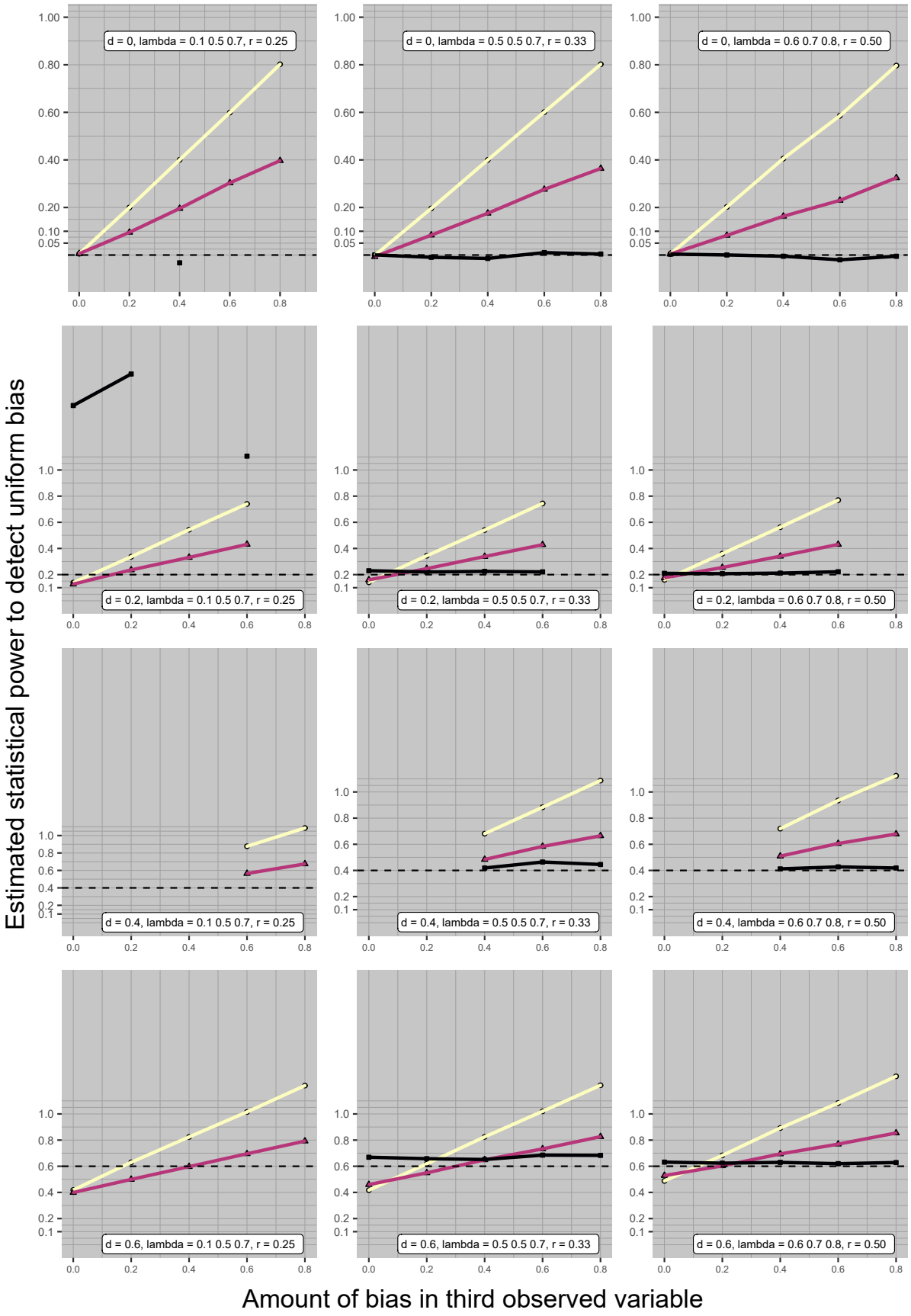

Figure 8: Estimated mean difference between groups for MGCFA, Sum ANOVA, and ANOVA 3 (biased) under uniform bias, for differing effect sizes (d). The dashed line indicates the true effect size.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance*. Beverly Hills: Sage Publications.

Breitsohl, H. (2019). Beyond ANOVA: An Introduction to Structural Equation Models for Experimental Designs. *Organizational Research Methods*, *22*(3), 649–677. https://doi.org/10.1177/1094428118754988

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Camilli, G., & Shepard, L. A. (1987). The Inadequacy of ANOVA for Detecting Test Bias. *Journal of Educational Statistics*, *12*(1), 87–99. https://doi.org/10.3102/10769986012001087

De Beuckelaer, A., & Swinnen, G. (2011). Biased Latent Variable Mean Comparisons Due to Measurement Noninvariance. In *Cross-cultural analysis: Methods and applications* (pp. 117–148). New York, NY, USA: Taylor & Francis Group.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I Error and Power of Latent Mean Methods and MANOVA in Factorially Invariant and Noninvariant Latent Variable Systems. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 534–556. https://doi.org/10.1207/S15328007SEM0704_2

Hartgerink, C. H. J., Wicherts, J. M., & Van Assen, M. A. L. M. (2017). Too Good to be False: Nonsignificant Results Revisited. *Collabra: Psychology*, *3*(1), 9. https://doi.org/10.1525/collabra.71

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). SAGE Publications, Inc.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed). New York: Guilford Press.

Meade, A. W., & Bauer, D. J. (2007). Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635. https://doi.org/10.1080/10705510701575461

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Revelle, W. (2020). *Psych: Procedures for Psychological, Psychometric, and Personality Research*.

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling, *48*(2), 1–36.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sörbom, D. (1974). A General Method for Studying Differences in Factor Means and Factor Structure Between Groups. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Steiger, J. H. (2016). Notes on the SteigerLind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 777–781. https://doi.org/10.1080/10705511.2016.1217487

Stevens, J. (2009). *Applied multivariate statistics for the social sciences* (5th ed). New York: Routledge.

Troncoso Skidmore, S., & Thompson, B. (2010). Statistical Techniques Used in Published Articles: A Historical Review of Reviews. *Educational and Psychological Measurement*, *70*(5), 777–795. https://doi.org/10.1177/0013164410379320

van Smeden, M., & Hessen, D. J. (2013). Testing for Two-Way Interactions in the Multigroup Common Factor Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 98–107. https://doi.org/10.1080/10705511.2013.742390

Venables, W. N., Ripley, B. D., & Venables, W. N. (2002). *Modern applied statistics with S* (4th ed). New York: Springer.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716. https://doi.org/10.1037/0022-3514.89.5.696

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37. https://doi.org/10.1037/1082-989X.8.1.16