FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

Factor Score Regression

Ines Devlieger

Supervisor: Prof. Dr. Yves Rosseel
Co-supervisor: Prof. Dr. Johan van Braak

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Educational Sciences

Academic year 2019–2020

GHENT
UNIVERSITY

Table of Contents

A sincere thank you ...
Een welgemeende dank u ...

... aan mijn promotor

Yves, ik kan niet anders dan dit dankwoord te beginnen bij jou. Toen ik zes jaar geleden begon, wist ik bijna niets over statistiek en data-analyse en nog minder over SEM en programmeren. Je nam de tijd om alles tot in het kleinste detail en op mijn niveau uit te leggen. Doorheen de jaren groeide mijn kennis en zelfvertrouwen en uiteindelijk kwamen we tot deze thesis. Ik had nooit gedacht dat dit het eindresultaat zou zijn (en ik weet zeker dat jij dit ook niet gedacht had). Ik kan je niet genoeg bedanken om mij de kans te geven om mezelf op mijn eigen tempo bij te scholen en te ontplooien. Ik ben dankbaar voor de kennis die je mij gegeven hebt, maar eigenlijk ben ik nog dankbaarder dat je me de kans gegeven hebt om een gezonde balans te vinden tussen mijn doctoraat en privé-leven. Ik ben nog nooit zo nerveus geweest als de dag dat ik moest komen vertellen dat ik zwanger was (dit is voor mijn verdediging geschreven, dus ik kan hierover nog van mening veranderen), maar jouw reactie was beter dan ik me ooit had kunnen voorstellen. Ik wil jou dan ook oprecht bedanken om me nooit het gevoel te geven dat ik moest kiezen tussen mijn doctoraat en mijn gezin.

... to my co-supervisor and members of the guidance committee

Johan, bedankt om mijn co-promotor te willen zijn, en Hilde, bedankt om deel te willen zijn van mijn begeleidingscommissie. Ik weet dat mijn doctoraat veel technischer is geworden dan we voor ogen hadden. Mijn excuses hiervoor. Ik wil jullie bedanken om steeds de moeite te doen om alles te volgen en vooral om er op toe te zien dat we de praktische kant van het verhaal niet uit het oog verloren. Axel, thank you

for your help during my first steps in the academic world and your sincere interest in my work later on. I learned a lot from our collaboration for the first paper of this thesis. Myrsini, thank you for your constructive feedback. I always enjoyed our talks.

... aan mijn collega's

Bedankt om een werkplek te creëren waar ik elke dag met veel plezier naar toe ben gegaan. Zelf tijdens de drukke periodes was er altijd een ontspannen sfeer en stond iedereen klaar om elkaar te helpen. Bedankt voor de uitgebreide theepauzes en de leuke vakgroepactiviteiten. Lara, Fien en Justine, ik heb genoten van onze tijd samen in bureau 17, al hebben de vele stekken naar elkaar me flink wat centjes gekost. Freya, wat zouden de eerste jaren van mijn doctoraat een stuk saaier geweest zijn zonder jou. Bedankt voor de vele gesprekken, face-to-face en online (Slack kan er van meespreken). Bedankt om zo mee te leven met de gebeurtenissen in mijn leven en om jouw zotte avonturen met mij te delen. Jasper, al op dag 1 stond je voor me klaar. Hoewel je zelf nog maar net begonnen was op de vakgroep, stond je extra vroeg op het werk om mij te kunnen ontvangen (pas later heb ik begrepen hoe speciaal dat was, toch in die tijd). En de volgende zes jaar kon ik altijd op je rekenen, of het nu was om me te helpen met de hpc, met werkgerelateerde problemen of gewoon voor een ontspannend gesprek. Heidelinde, laat ons eerlijk zijn, wij namen niet de beste start als collega's. Maar eens we samen in een bureau terecht kwamen, verdwenen de spanningen al snel. Bedankt om de ritjes naar het verre West-Vlaanderen af en toe op te leuken met jouw aanwezigheid. Wouter, we bleven lang verre collega's. Na de verhuis belandden we echter op dezelfde bureau en bleken onze kinderen al snel een goed bindmiddel te zijn, ondanks het feit dat ze complete tegenpolen zijn. Bedankt voor de vele interessante discussies, of het nu over opvoeding, werk of levenskwesties ging. Ik ben

heel blij dat je mee geweest bent naar Jena en nog enkele maanden langer bent blijven plakken op de vakgroep om ons uit de nood te helpen. Bieke, bedankt om me te steunen als jonge moeder, om verhalen over jouw kroost te delen en om te luisteren naar de anekdotes over mijn kroost. Je bent een fantastisch rolmodel voor vrouwen in de academische wereld. Jan, bedankt om mij te vertrouwen met jouw vak en me zoveel verantwoordelijkheden te geven. Ik heb er van genoten om af en toe een onderwijskundig experimentje te kunnen uitvoeren, zodat mijn vooropleiding niet volledig in de vergetelheid verdween. Isabelle, bedankt dat jouw deur altijd voor mij (en de rest van de vakgroep) open stond en niet alleen voor administratieve kwesties. Ik zal mijn pauzes in jouw bureau missen, al zal jij waarschijnlijk een stuk meer rust hebben zonder mij in de buurt. Bedankt voor het speelhuisje, de werkbank en alle andere cadeautjes die mijn kinderen van jou gekregen hebben.

... aan mijn familie

Bedankt om te vormen tot wie ik ben. Zonder jullie was ik nooit aan een doctoraat kunnen beginnen. Mama en papa, bedankt dat jullie me altijd alle kansen hebben gegeven om te doen wat ik zelf wou. Jullie hebben me nooit een bepaalde richting ingeduwd, zelfs niet toen ik na één week universiteit al van opleiding wilde veranderen. Pepe Freddy en meme Josée, bedankt voor de leuke tijden die ik bij jullie doorgebracht heb, zowel na school thuis als tijdens de zomer aan zee. Pepe, wat had ik graag gehad dat je er nog bij was om te zien dat we allemaal goed terecht gekomen zijn. Pepe Gerard en meme Yvonne, bedankt voor de leuke uitstapjes en reizen en bedankt om zo goed voor de kindjes te zorgen op donderdag. Kirsten en Ellen, bedankt voor de zorgeloze jeugd samen, met geregeld wat kattekwaad waarvan af en toe anekdotes komen bovendrijven tijdens familie-etentjes. Bedankt om de beste petie en metie te zijn die Matteo en Mila zich kunnen inbeelden. Wat is het fantastisch om onze kinderen nu ook samen te

zien opgroeien. Elien, Delphine, Bert en Elien, Sander en Lore, bedankt voor de leuke familiemomenten, bedankt voor de prachtige nichtjes en het schattige neefje en bedankt om zulke leuke tantes en nonkels te zijn. Noor en Suzan, Iliana, Ilaura en Iluca, Lina, wat is het een voorrecht om jullie tante (en metie) te mogen zijn. Stuk voor stuk prachtige mensjes, met een eigen karaktertje en eigen willetje. Blijf alsjeblieft altijd jezelf. Mama, papa, Lut en Guido, bedankt om zulke fantastische en actief betrokken grootouders te zijn. Jullie staan altijd klaar om in te springen. Het combineren van mijn doctoraat en gezin was een stuk lastiger geweest zonder jullie hulp.

... aan mijn gezin

Jeroen, wat is er veel veranderd de laatste zes jaar. We kochten een huis, verbouwden het volledig, trouwden en kregen twee prachtige kinderen. Ons leven stond geen seconde stil, maar onze relatie bleef altijd stabiel. De stabiliteit in ons privé-leven gaf me de nodige rust om me op mijn doctoraat te kunnen focussen. Je zorgde er ook voor dat ik mijn doctoraat nooit te serieus nam. Matteo en Mila, bedankt dat ik jullie mama mag zijn. Jullie veranderden mijn kijk op de wereld en op wat belangrijk is. Jullie plaatsten alles in perspectief: ik ben trots op mijn doctoraat, maar ik ben nog veel trotser op ons prachtig gezin. Ik kijk vol verwachting uit naar wat de toekomst ons nog zal brengen.

Ines,
Oktober 2019

1 Introduction

In this dissertation, I address one of the main methodological challenges educational sciences deals with. To explore this challenge, I list below some of the research papers educational researchers have published in the last couple of years:

- Transformational school leadership as a key factor for teachers' job attitudes during their first year in the profession (Thomas, Tuytens, Devos, Kelchtermans, & Vanderlinde, 2018)
- Identifying student and classroom characteristics related to primary school students' listening skills : a systematic review (Bourdeaud'hui, Aesaert, Van Keer, & van Braak, 2018)
- Effects of immersion in inquiry-based learning on student teachers' educational beliefs (Voet & De Wever, 2018)
- The influence of motivation and comfort-level on learning to program (Bergin & Reilly, 2005)

Many of the concepts that are studied in these papers, e.g school leadership, job attitude, listening skills, educational beliefs, and motivation, are very hard or even impossible to measure directly. Such variables are referred to as latent variables. I will use the concepts from the last paper as a running example throughout this introduction. This example is depicted in Figure 1.1. How does one measure motivation or programming skills? And how do you study the influence of motivation on programming skills? This will be the core of this dissertation: how to study relationships between latent

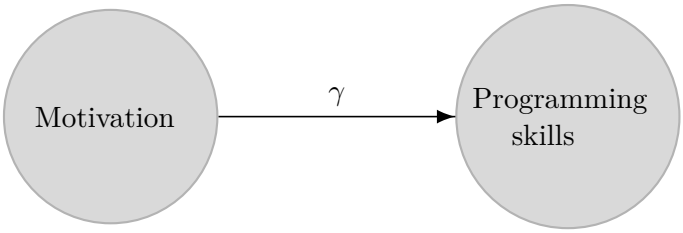


Figure 1.1 Example: the relationship between motivation and programming skills

variables. I am thus mainly interested in the regression parameter γ, which can be seen in Figure 1.1. In this introduction, I first describe how to measure latent variables, before describing how to study the relationships between latent variables.

1.1 How to measure latent variables

Latent variables cannot be measured directly, so they are usually measured by using observed indicators. Motivation, for example, is often measured by using questionnaires that contain multiple questions. These questions will be referred to as items. Afterwards, the information from the different items has to be combined into one construct. This can be done by using factor analysis (FA).

1.1.1 The model

Assuming $\mathbf{y}$ is mean-centered, the measurement model that is used for a factor analysis with a single factor is the following:

$$\mathbf{y} = \boldsymbol{\lambda}\eta + \boldsymbol{\varepsilon}. \tag{1.1}$$

Denote $var(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$, $var(\eta) = \boldsymbol{\sigma}_{\eta}^2$ and assume $cov(\eta, \boldsymbol{\varepsilon}) = \mathbf{0}$. There are four elements in this equation:

- η represents the latent variable that one is trying to measure. In our example, this is 'motivation'.

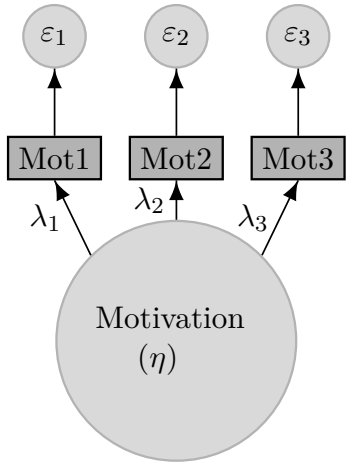


Figure 1.2 The measurement model for the latent variables 'motivation'.

- $\mathbf{y}$ is a vector of the observed items that measure the latent variable η. The variance-covariance matrix of $\mathbf{y}$ is denoted as $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \boldsymbol{\sigma}_{\eta}^{\mathbf{2}} \boldsymbol{\lambda} \boldsymbol{\lambda}^T + \mathbf{\Theta}$$

- $\boldsymbol{\lambda}$ is a vector containing the factor loadings, which represent the linear relationships between the items and the latent variable.

- $\boldsymbol{\varepsilon}$ is a vector containing the residual errors. This is the error with which the items are measured. All items are measured with a certain error, meaning they do not measure the true latent variable exactly. If there was no error, the relationship between the item indicator and the true latent variable would be perfect, implying there is no need to measure multiple item indicators.

This model is depicted in Figure 1.2 for our example of motivation. To be able to estimate all the parameters in this model, the metric scale of the latent variable has to be fixed. This can be done by fixing one of the factor loadings to 1 or by fixing the variance of the

latent variable to 1.

1.1.2 Estimation

Different methods can be used to estimate the free parameters in the factor analysis model, such as Maximum Likelihood Estimation (MLE), Weighted Least Squares (WLS) and Generalized Least Squares (GLS). When the items are continuous, MLE is the most commonly used method (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Using MLE, the parameters are estimated by minimizing the discrepancy function

$$\mathrm{F}_{ML} = tr(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}) + log|\hat{\mathbf{\Sigma}}| - log|\mathbf{S}| - k, \tag{1.2}$$

with $\mathbf{S}$ the observed variance-covariance matrix, $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}(\boldsymbol{\theta})$ the model implied variance-covariance matrix, given the vector of free parameters $\hat{\boldsymbol{\theta}}$, and k the number of observed variables. In other words, the values of the parameters in $\boldsymbol{\theta}$ are chosen in such a way that the difference between the observed variance-covariance matrix $\mathbf{S}$ and the model implied variance-covariance matrix $\hat{\mathbf{\Sigma}}$ is as small as possible.

1.1.3 Factor scores

Once all the parameters are estimated, scores for the latent variable can be calculated by multiplying a factor score matrix $\mathbf{A}$ with the observed indicators $\mathbf{y}$:

$$\mathbf{F} = \mathbf{A}\mathbf{y}. \tag{1.3}$$

The resulting scores are referred to as factor scores. The computation of the factor score matrix $\mathbf{A}$ depends on the predictor that is used. There are several predictors and all of them incorporate both the factor loadings and the residual errors into the factor score matrix one way or another. The most commonly used predictors are the Regression predictor (Thomson, 1934; Thurstone, 1935) and the Bartlett predictor (Bartlett, 1937; Thomson, 1938). A more detailed description of these predictors will be given in Chapter 2.

1.1.4 Scale scores

Another, often used, way to obtain scores for the latent variable, is to simply add up the scores of all indicator items or take the average. These are referred to as scale scores. The problem with this approach is that it relies on two underlying assumptions that are usually violated in reality. First, it assigns equal weights to all items. This implies that all items are measured on exactly the same scale. So, every question on the motivation questionnaire has exactly the same relationship with your true motivation. Secondly, scale scores assume that the residual errors of the observed item indicators are all equal (Graham, 2006). This means that the reliability is the same for every item. Both of these assumptions are unlikely to be true, implying that scale scores are not an accurate estimate of your true motivation. Note that if you constrain the factor loadings and residual errors of a FA to be equal, they are equivalent to scale scores up to a linear transformation.

1.2 How to study the relationship between the latent variables

Most often researchers are not really interested in measuring the latent variable itself. They are mainly interested in the relationships between the latent variables and the relationships between latent variables and observed variables. For simplicity reasons, I will refer to these relationships as regression parameters. Trying to get a reliable, unbiased estimate for these regression parameters is the core of this dissertation.

1.2.1 Structural Equation Modeling

Structural Equation modeling (SEM) estimates the parameters of the measurement model and the regression parameters in one step (Bollen, 1989). SEM consists of two models, namely the measurement model

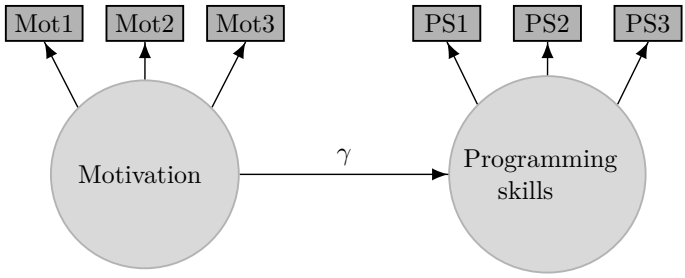


Figure 1.3 Using structural equation modeling to estimate the relationship between motivation and programming skills. Note that the residual errors were left out of the figure. This is for simplicity reasons and will be done in all the figures hereafter.

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1.4}$$

and the structural model

$$\boldsymbol{\eta} = \boldsymbol{\gamma}\boldsymbol{\eta} + \boldsymbol{\zeta}, \tag{1.5}$$

assuming $cov(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \mathbf{0}$. Different methods can be used to estimate the parameters, but in general MLE is used when the items are continuous. When I refer to SEM in this dissertation, I assume MLE is used. SEM also minimizes the discrepancy function 1.2, but the vector of free parameters $\boldsymbol{\theta}$ now also includes the parameters from the structural model. A diagram of the model for the influence of motivation on programming skills can be found in Figure 1.3.

SEM has several advantages. Firstly, SEM takes the measurement error into account when estimating the regression parameters, leading to unbiased results. Secondly, fit measures can be obtained that give an indication of how well the model fits the data. Most fit measures are based on the the χ^2-test statistic. This test statistic is based on the difference between the covariance matrix of the observed data ($\mathbf{S}$) and the model-implied covariance matrix ($\hat{\boldsymbol{\Sigma}}$).

However, the simultaneous estimation of all parameters also has some downsides. Because of the large number of parameters that

have to be estimated at once, a large sample size is needed to obtain these unbiased results. In practice, it is often hard to obtain a sample size that is sufficiently large to perform SEM, due to time and financial restrictions. Also, if part of the model is misspecified, it can potentially affect all parameters. In conclusion, SEM is considered to be the 'gold standard', but it requires a correctly specified model and a rather large sample size. This makes it rather impractical for applied researchers.

1.2.2 Naive factor score regression

Applied researchers often turn to a stepwise method, where they first calculate the factor scores and then simply use them in a subsequent analysis such as linear regression, mediation analysis or multilevel analysis. This is refered to as factor score regression (FSR). This procedure has been depicted for our example of the influence of motivation on programming skills in Figure 1.4.
When doing this, one actually treats the factor scores as if they are observed scores. The problem with this approach is that this is simply not true. In fact, as described above, there are several ways to compute factor scores, all resulting in different predicted factor scores that are all equally viable. This is referred to as factor indeterminacy (Maraun, 1996; Mulaik, 1972; Steiger, 1979). The higher the indeterminacy, the larger the differences in the factor scores of the different predictors. When the indeterminacy is low, the different predictors will lead to largely the same factor scores. The indeterminacy will typically be low if the relationship between the latent variable and the items is strong and will be high if the relationship is weak. Or in other words: the better the item indicators measure the latent variable, the better the latent variables scores can be predicted. However, the factor scores will still be a prediction and not an observation. Factor indeterminacy implies that it is impossible to obtain an unambiguous prediction for the true latent variable scores (Lastovicka & Thamodaran, 1991). There is always a certain degree of uncertainty. As a consequence, the co-

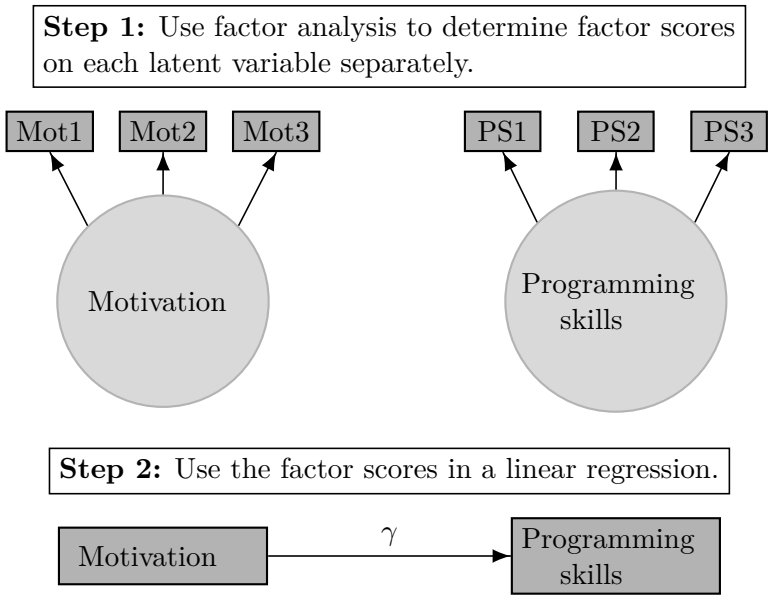


Figure 1.4 Using factor score regression to estimate the relationship between motivation and programming skills

variance matrix of the factor scores is never completely the same as the covariance matrix of the true latent variable scores:

$$var(\mathbf{F}) \neq var(\boldsymbol{\eta}) \tag{1.6}$$

This is completely disregarded when the factor scores are used in a subsequent analysis, and will often lead to results that are biased. In chapter 2, an overview of the settings where this leads to bias is given.

1.2.3 Correction methods

Several researchers have tried to develop a stepwise method, that also avoids or corrects for the bias that is inherent to FSR.

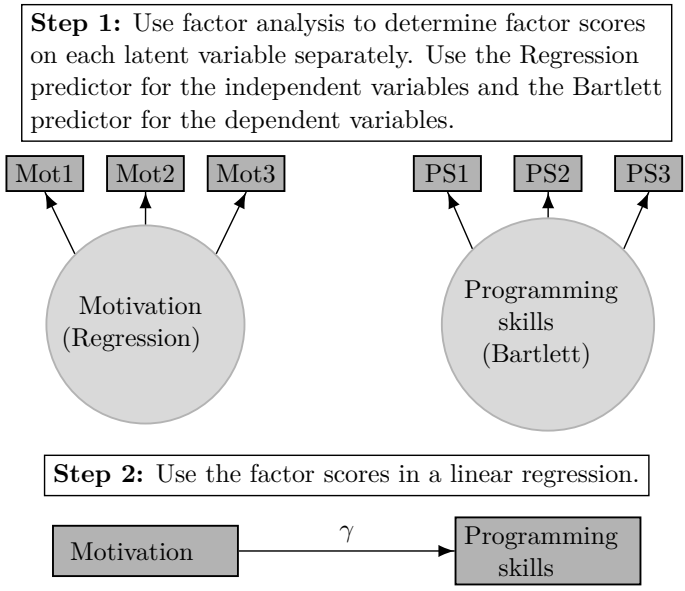


Figure 1.5 Using factor score regression with bias avoiding to estimate the relationship between motivation and programming skills

Bias avoiding

Skrondal and Laake (2001) tried to avoid the bias by using different predictors for the independent and dependent latent variables. As depicted in Figure 1.5, the Regression predictor is used for the independent latent variables and the Bartlett predictor is used for the dependent latent variables. This results in an unbiased estimate for γ.

Bias correcting

A different tactic is to correct for the bias by correcting the variance-covariance matrix of the factor scores. The regression parameters in a simple linear regression can be calculated by dividing the covariance between the dependent and independent variable by the variance of the independent variable. In our example, γ can be

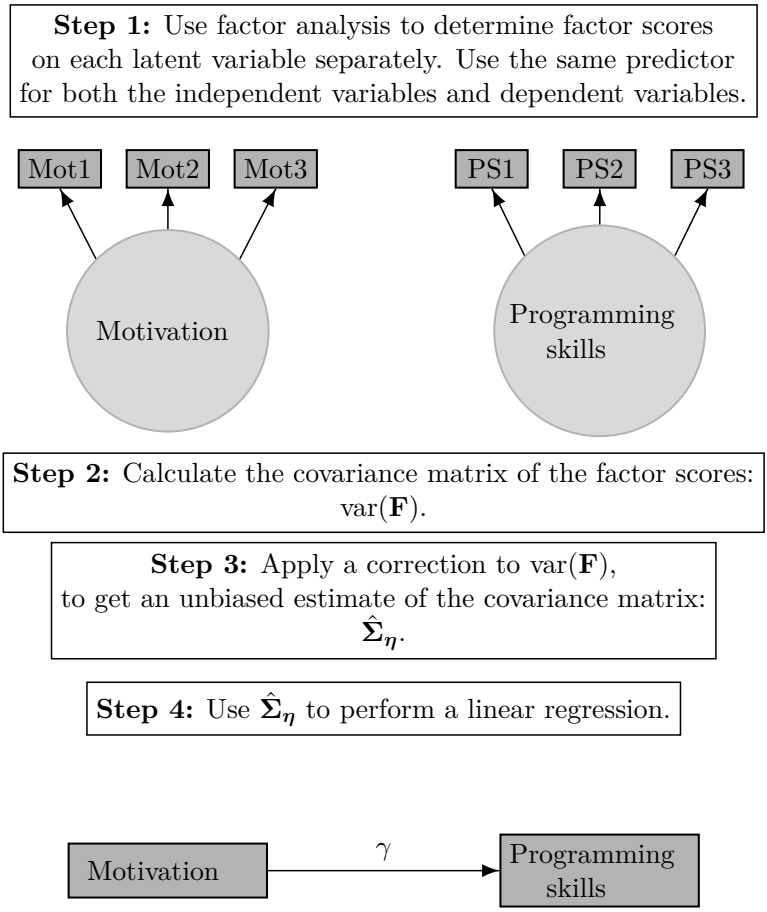


Figure 1.6 Using factor score regression with bias correcting to estimate the relationship between motivation and programming skills

calculated as follows:

$$\gamma = \frac{cov(motivation, programming)}{var(motivation)}. \tag{1.7}$$

This means that the raw scores are not needed to perform a linear regression: having a variance-covariance matrix is sufficient. A corrected variance-covariance matrix of the factor scores can thus be used to calculate the regression parameters. An overview of this strategy can be found in Figure 1.6. There are several methods that do this, including methods suggested by Croon (2002) and Hoshino and Bentler (2013). The exact way these methods correct the variance-covariance matrix will be discussed in further detail in chapter 2 and 3. While these methods are very promising, they remain rather unknown and understudied. The settings that they are studied in are rather limited, usually using a very simple linear regression model, so many questions remain unanswered:

- Which of these methods performs the best?
- Can these methods indeed overcome the limitations of SEM with regard to sample size and model misspecifications?
- Do these methods also work in other settings, such as mediational models and multilevel models?
- How does one assess the model fit when using stepwise methods?
- How to estimate the standard error when using stepwise methods?

1.3 Objectives

In this dissertation, I will try to give an answer to these questions. In general, I have three main objectives, namely comparing methods, expanding stepwise methods to other settings than linear regression and performing hypothesis and model testing.

1.3.1 Comparing methods

First of all, it is important to know which of the several stepwise methods performs the best. In chapter 2, a comparison of naive factor score regression, one bias avoiding method and one bias correcting method is given. The methods are compared by means of a simulation study, using several performance criteria such as convergence rate, bias in the unstandardised setting, bias in the standardised setting, efficiency, MSE, power and type I-error rate. At the end of this chapter, it is concluded that the bias correcting method suggested by Croon (2002) is the best stepwise method.
Next to knowing wich stepwise method is best, it is also important to now how it compares to the 'gold standard' SEM. In chapter 2, SEM is also included in the simulation study to see if the stepwise methods can perform equally well as SEM in a linear regression setting. In chapter 3, SEM is compared to the method of Croon to see if the method of Croon can even outperform SEM, especially with regard to the main limitations of SEM, namely small sample sizes and misspecifications.

1.3.2 Expanding the settings

Mediation

In chapter 3, the method of Croon is implemented in mediational settings. In mediational settings, the influence of one variable on another can be both direct and indirect through another variable. For example, the influence of motivation on programming skill could be mediated by the amount of time a person spends practicing. This is depicted in Figure 1.7. So, motivation has a direct influence on the programming skills, but it also has an influence on the hours of practice, which in turn also influences the programming skills.

Multilevel regression

In chapter 5, the method of Croon is extended to the multilevel setting. In educational sciences, the data is often hierarchical in

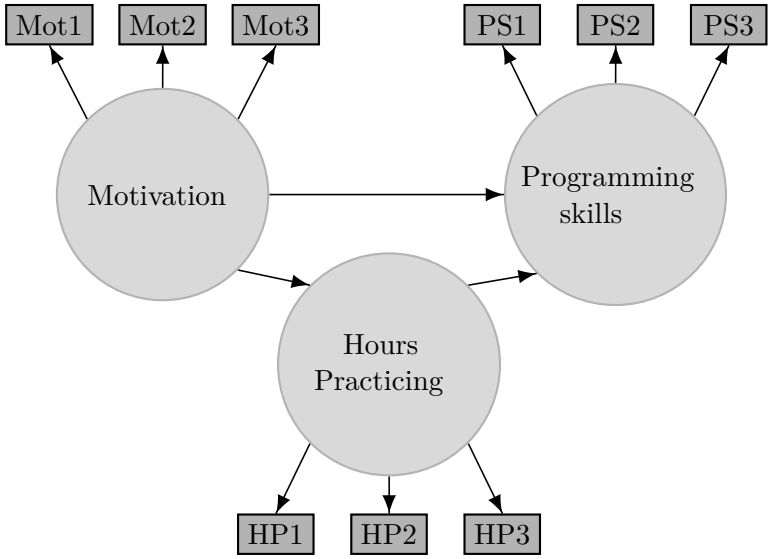


Figure 1.7 A mediational model: the influence of 'motivation' on 'programming skills' is mediated by 'hours practicing'.

nature. For example, if one studies pupils, these pupils are usually members of a school. Pupils from the same school tend to be more alike compared to pupils from different schools. In other words, pupils are clustered within classes and classes are clustered within schools. In our motivation example, the pupils can be members of different schools (see Figure 1.8). This means that not all pupils in our dataset are independent from each other. However, this is an assumption of single-level linear regression and mediation. The clustered nature of the data requires specialized techniques.

1.3.3 Performing hypothesis and model testing

Being able to estimate unbiased regression parameters is very important, but also insufficient. To be able to draw statistical inference, a reliable standard error is necessary. The stepwise nature of the Croon method makes this hard to obtain. In chapter 2, an ad-hoc solution for the linear regression setting is proposed. However, this solution was not transferable to the mediational and multilevel

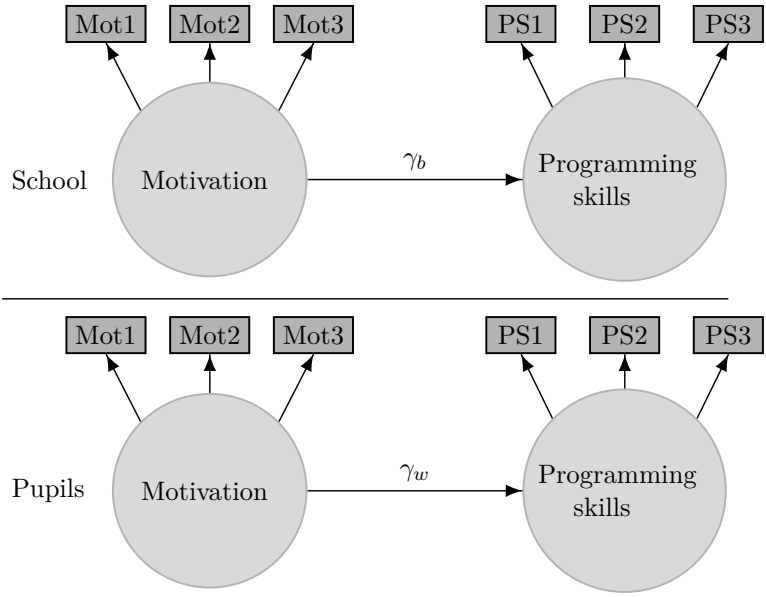


Figure 1.8 Multilevel model to estimate the relationship between motivation and programming skills

setting. In chapter 3, the use of bootstrapping to estimate the standard errors is proposed. In chapter 5, an analytical approach that was developed by Bakk, Oberski, and Vermunt (2014) is used. However, this approach is mainly useful to test single parameters. It cannot be used to perform a model comparison, nor can it be used to evaluate the fit of the model. In chapter 4, an approximate χ^2-test-statistic is introduced that can be used to construct fit indices and to perform model comparison tests.

1.4 Overview of the chapters

Chapter 2 In this chapter, several methods in a linear regression context are compared, to be able to determine which stepwise method is to be preferred. At the end of this chapter, it is conclude that the method of Croon is the best of the stepwise methods. The chapter is published as

> Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76, 741–770. doi: 10.1177/0013164415607618

Chapter 3 In the third chapter, the method of Croon is expanded to the mediational setting. Moreover, it is demonstrated that the method of Croon can outperform SEM in settings with a complex model and/or a small sample size. The chapter is published as

> Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology*, 13, 31–38. doi: 10.1027/1614-2241/a000130

Chapter 4 In the fourth chapter, a way to calculate an approximate χ^2-test statistic for the full model is developed. This χ^2-test statistic can be used to perform model comparison tests and to calculate fit indices to assess how well the model fits the data. The chapter is published as

> Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New Developments in Factor Score Regression: Fit Indices and a Model Comparison Test. *Educational and Psychological Measurement*, 79(6), 1017–1037 doi: 10.1177/0013164419844552

Chapter 5 In the fifth chapter the method of Croon is expanded to the multilevel setting. The chapter is published as

> Devlieger, I., & Rosseel, Y. (2019). Multilevel factor score regression. *Multivariate Behavioral Research*, doi: 10.1080/00273171.2019.1661817

Chapter 6 In this chapter, I provide a general discussion of the findings in this dissertation. I also discuss the limitation and perspectives for future research.

Chapter 7 In this chapter I provide a summary of this dissertation.

In chapters 2 to 5 there might be some overlap since these chapters have been written as articles. As a consequence I introduce for each chapter notation, which might not be identical across the chapters.

References

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014, jan). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, *22*(4), 520–540. doi: 10.1093/pan/mpu003

Bartlett, M. (1937, jul). The statistical conception of mental factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Bergin, S., & Reilly, R. (2005). The influence of motivation and comfort-level on learning to program. *Ppig 17*(June), 293–304. doi: 10.1.1.443.427

Bollen, K. (1989). *Structural equations with latent variabels*. New York: NY: Wiley.

Bourdeaud'hui, H., Aesaert, K., Van Keer, H., & van Braak, J. (2018). Identifying student and classroom characteristics related to primary school students' listening skills: A systematic review. *Educational Research Review*, *25*(September), 86–99. doi: 10.1016/j.edurev.2018.09.005

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. *ducational andPsychological Measurement*, *66*(6), 930–944.

Hoshino, T., & Bentler, P. M. (2013). Bias in Factor Score Regression and a Simple Solution. In A. R. de Leon & K. C. Chough (Eds.), *Analysis of mixed data- methods & applications* (pp.

43–61). Chapman and Hall. doi: 10.1201/b14571-5

Lastovicka, J. L., & Thamodaran, K. (1991). Common factor score estimates in multiple regression problems. *Journal of Marketing Research*, *28*(1), 105–112. doi: 10.2307/3172730

Maraun, M. D. (1996). Metaphor taken as math: Indeterminancy in the factor analysis model. *Multivariate Behavioral Research*, *31*(4), 517–538. doi: 10.1207/s15327906mbr3104_6

Mulaik, S. A. (1972). Factor scores and factor indeterminacy. In S. A. Mulaik (Ed.), *Foundations of factor analysis.* New York: McGraw-Hill Book Company.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74. doi: 10.1002/0470010940

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. doi: 10.1007/BF02296196

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, *44*(2), 157–167. doi: 10.1007/BF02293967

Thomas, L., Tuytens, M., Devos, G., Kelchtermans, G., & Vanderlinde, R. (2018). Transformational school leadership as a key factor for teachers' job attitudes during their first year in the profession. *Educational Management Administration and Leadership*. doi: 10.1177/1741143218781064

Thomson, G. (1934). The meaning of i in the estimate of g. *British Journal of Psychology*, *25*, 92–99. doi: 10.1111/j.2044-8295.1934.tb00728.x

Thomson, G. (1938, feb). Methods of Estimating Mental Factors. *Nature*, *141*(3562), 246–246. doi: 10.1038/141246a0

Thurstone, L. (1935). *The vectors of mind.* Chicago, IL: University of Chicago Press. doi: 10.1037/h0075959

Voet, M., & De Wever, B. (2018). Effects of immersion in inquiry-based learning on student teachers' educational be-

liefs. *Instructional Science*, *46*(3), 383–403. doi: 10.1007/s11251-017-9439-8

2 Hypothesis testing using factor score regression: A comparison of four methods

Abstract. In this article, an overview is given of four methods to perform Factor Score Regression (FSR), namely regression FSR, Bartlett FSR, the bias avoiding method of Skrondal and Laake (2001) and the bias correcting method of Croon (2002). The bias correcting method is extended to include a reliable standard error. The four methods are compared to each other and to SEM by using analytic calculations and two Monte Carlo simulation studies to examine their finite sample characteristics. Several performance criteria are used, such as the bias using the unstandardized and standardized parameterization, efficiency, mean square error, standard error bias, type I-error rate and power. The results show that the bias correcting method, with the newly developed standard error, is the only suitable alternative for SEM. While it has a higher standard error bias than SEM, it has a comparable bias, efficiency, MSE, power and type I-error rate.

This chapter has been published as Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76, 741–770. doi: 10.1177/0013164415607618.

2.1 Introduction

In the social and behavioral sciences, the aim of applied researchers is often to examine the relationships between latent variables. Latent variables are variables that are not directly observable, such as intelligence, skill or motivation. To measure these latent variables, observable indicators are used (Bollen & Hoyle, 2012). Structural Equation Modeling (SEM) can be used to simultaneously and consistently estimate both the measurement models and the structural relations between these latent variables (Bentler & Chou, 1987; Jöreskog, 1973).

Despite the increasing popularity of SEM, many applied researchers prefer to use the Factor Score Regression (FSR) method, which is more intuitive and consists of two steps. In a first step, the scores on the latent variables are predicted using factor analysis (FA). In this paper, we will refer to these predicted scores as factor scores. In a second step, the factor scores are used in a linear regression (OLS) (Lu, Thomas, & Zumbo, 2005). Unfortunately, there are an infinite number of ways to compute these factor scores, all of which are consistent with the FA performed (Grice, 2001), meaning they are all equally viable. The two most commonly used predictors are the regression predictor (Thomson, 1934; Thurstone, 1935) and the Bartlett predictor (Bartlett, 1937; Thomson, 1938). The factor scores will be different depending on which predictor is used.

This phenomenon is referred to as factor indeterminacy (Maraun, 1996; Mulaik, 1972; Steiger, 1979). The degree of indeterminacy is small if the relationship between the indicators and the latent variable is strong or if the number of indicators is high (Acito & Anderson, 1986). When there is a high degree of factor indeterminacy, it is even possible for an individual to score high according to the factor scores calculated by one method and score low according to the factor scores of another method (Grice, 2001). Lastovicka and Thamodaran (1991) conclude that indeterminacy implies that factor scores cannot be measured and researchers have to accept that factor analysis means it is impossible to obtain an unambigu-

ous prediction or computation of the latent variable scores. In other words, a degree of uncertainty is inherent to factor scores, which is no longer accounted for when using factor scores in linear regression. This uncertainty causes the regression coefficient to be biased, as has been discussed extensively in the literature (e.g. Bollen, 1989; Lastovicka & Thamodaran, 1991; Lewis, 2005; Shevlin, Miles, & Bunting, 1997). Despite its obvious drawbacks, FSR remains a popular method among applied researchers (Lu & Thomas, 2008).

For this reason, improved methods to perform FSR have been developed, which result in an unbiased regression coefficient (Croon, 2002; Skrondal & Laake, 2001). Skrondal and Laake (2001) developed a method which avoids bias altogether, while Croon (2002) found a method to correct for the bias. We will refer to these methods as the bias avoiding and bias correcting method, respectively. Both methods indeed result in an unbiased parameter estimate, but are hardly ever used in practice. This is partly due to the highly technical and mathematical level of the papers describing the methods. Lu, Kwan, Thomas, and Cedzynski (2011) tried to remedy this by giving an overview of both methods. The statistical performance of both methods was also compared in a simulation study, with regard to accuracy and power. However, the methods remain unused by applied researchers. This could be due to some practical issues. First of all, it is only described how to obtain an unbiased estimate for the regression coefficient, but for the bias correcting method, there is no standard error available. This means that this method cannot be used yet to test hypotheses. Second, the results of the methods have only been described for the unstandardized parameterization. Neither Skrondal and Laake (2001), Croon (2002) or Lu et al. (2011) describe what happens when the standardized parameterization is used. In conclusion, the methods are not directly usable for the applied researchers.

The goal of this paper is to compare various methods for performing FSR, namely FSR using the regression predictor (regression FSR), FSR using the Bartlett predictor (Bartlett FSR), the method of

Skrondal and Laake (2001) (bias avoiding method) and the method of Croon (2002) (bias correcting method). For each method we derive the bias analytically on the population level for both the standardized and unstandardized parameterization. Next, for the bias correcting method, a new standard error is developed, making it possible to use the bias correcting method to test hypotheses.
Finally, two simulation studies are set up to compare the performance of the four methods in finite samples, using normal and non-normal data, respectively. The simulation studies also allow us to evaluate the performance of the newly developed standard error. Since SEM is generally considered as the standard method to examine the regressions between latent variables, SEM is also included in the simulation studies. The aim is to be able to compare the methods on their overall statistical performance and formulate recommendations for the applied users.

2.2 Setting

To be able to compare the four methods, a simple regression model with one dependent and one independent variable is used. The simple regression model is used to reduce the notational complexity and enhance comprehensibility. However, it can easily be extended to settings with more than one dependent variable and more than one independent variable. In fact, a more complex setting is used in the simulation study.
Within this framework of a simple linear regression, we consider four possible scenarios, which are visualized in Figure 2.1. In the first scenario, both the dependent and independent variable are measured without error. They are considered observed variables. In the second scenario, factor scores are used for the independent variable, while the dependent variable is observed. In the third scenario, factor scores are used for the dependent variable, while the independent variable is observed. In the fourth and final scenario, factor scores are used for both variables. In all scenarios, the

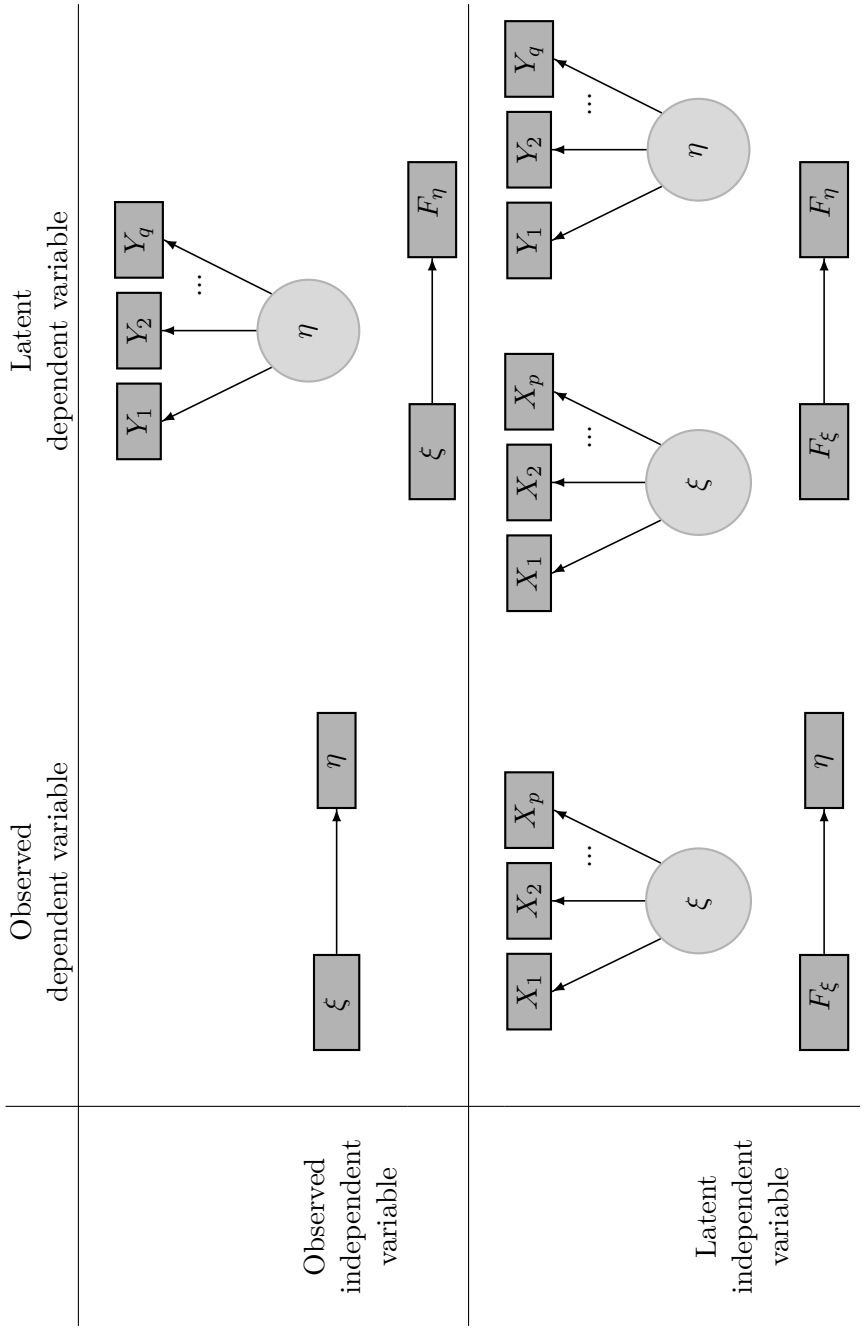


Figure 2.1 The four scenarios considered

structural equation is:

$$\eta = \gamma\ \xi + \zeta, \tag{2.1}$$

where η is the dependent variable, ξ is the independent variable, γ is the regression coefficient and ζ is the residual error term. When η and ξ are latent, the following measurement models are used

$$\mathbf{x} = \mathbf{\Lambda}_x \xi + \boldsymbol{\delta} \tag{2.2}$$

$$\mathbf{y} = \mathbf{\Lambda}_y \eta + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{x} = (X_1, ..., X_i, ..., X_p)^T$ and $\mathbf{y} = (Y_1, ..., Y_j, ..., Y_q)^T$ are vectors of mean-centered observed indicators measuring ξ and η respectively, $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ are vectors of the factor loadings and $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are the respective vectors of measurement error variables.

In the first step of factor score regression, we use these measurement models to perform a FA for each latent variable separately and to calculate the factor scores for ξ (F_ξ) and η (F_η). To be able to perform a FA, the metric scales of the latent variables ξ and η have to be fixed. This can de done in several ways, for example by fixing one factor loading per latent variable to 1 or by fixing the variance of the latent variable to 1. We will refer to the latter as the standardized parameterization, and the former as the unstandardized parameterization. In this paper, we will mainly use the unstandardized parameterization.

The factor scores are calculated by multiplying a factor score matrix $\mathbf{A}$ with the observed indicators $\mathbf{x}$ (or $\mathbf{y}$):

$$F_\eta = \mathbf{A}_\eta \mathbf{y} \tag{2.4}$$

$$F_\xi = \mathbf{A}_\xi \mathbf{x}. \tag{2.5}$$

The computation of the factor score matrices $\mathbf{A}_\eta$ and $\mathbf{A}_\xi$ depends on the method used for the prediction of the factor score. The different methods and their influence will be discussed in the next

section.
In the second step of factor score regression, a linear regression is performed between the factor scores, resulting in a regression coefficient. In a simple linear regression, the true regression coefficient is defined as the true covariance between the dependent and the independent variable, divided by the variance of the independent variable:

$$\gamma = \frac{cov(\xi, \eta)}{var(\xi)}. \tag{2.6}$$

When performing the linear regression with factor scores, the regression coefficient becomes

$$\beta = \frac{cov(F_\xi, F_\eta)}{var(F_\xi)}, \tag{2.7}$$

which is not necessarily the same as the true regression coefficient. The relationship between γ and β can best be understood if we work out the covariance and variances of the factor scores. In Appendix A1, we derive the exact relationship between β and γ:

$$\beta = \frac{cov(F_\xi, F_\eta)}{var(F_\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x var(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} \gamma. \tag{2.8}$$

From this, it is clear that in most cases the regression coefficient obtained with factor score regression will not be the same as the true regression coefficient. It is also possible to calculate the expected regression coefficients when factor scores are only used for one of the variables, dependent or independent. The interested reader can find the calculations in Appendices A2 and A3.

2.3 Methods to perform FSR

To perform FSR, several methods can be used, such as the regression FSR method, the Bartlett FSR method, the bias avoiding method (Skrondal & Laake, 2001) and the bias correcting method (Croon, 2002). In this section, these four methods will be discussed.

In Table 2.1, an overview of this discussion is given.

2.3.1 Regression FSR method

The regression FSR method uses the regression predictor (Thomson, 1934; Thurstone, 1935) to compute the factor scores and then uses these factor scores in a linear regression. When using the regression predictor, the factor scoring matrices $\mathbf{A}_{\xi}^{R}$ and $\mathbf{A}_{\eta}^{R}$ are computed as follows:

$$\mathbf{A}_{\xi}^{R} = \mathbf{\Phi}\mathbf{\Lambda}_{x}^{'}\mathbf{\Sigma}_{x}^{-1} = var(\xi)\mathbf{\Lambda}_{x}^{'}\mathbf{\Sigma}_{x}^{-1}, \tag{2.9}$$

and

$$\mathbf{A}_{\eta}^{R} = var(\eta)\mathbf{\Lambda}_{y}^{'}\mathbf{\Sigma}_{y}^{-1}. \tag{2.10}$$

This means the formula for the variance of F_{ξ} can be simplified. In appendix B, this is done to show that the regression coefficient in factor score regression is not biased in all cases. When using the regression FSR method, there is only bias when factor sores are used for the dependent variable. With regard to bias, it is acceptable to use the regression FSR method if only the independent variables are factor scores. When factor sores are used for the dependent variable, one should not use regression FSR, since the regression parameter will be biased in most cases.

These results only apply when the unstandardized parameterization is used (Skrondal & Laake, 2001). The standardized regression coefficient γ_z is different from the unstandardized regression coefficient γ. The standardized regression coefficient can be calculated from the unstandardized regression coefficient and standard deviations:

$$\gamma_z = \frac{sd(\xi)}{sd(\eta)}\gamma. \tag{2.11}$$

Since $sd(F_{\xi})$ and $sd(F_{\eta})$ are biased, the standardized regression coefficient will be biased too if factor scores are used for any of the variables.

Table 2.1 The regression parameter β in relationship to γ when using FSR

$\beta =$	both variables observed	independent variable factor scores, dependent variable observed	independent variable observed, dependent variable factor scores	both variables factor scores
The general case	γ	$\frac{\mathbf{A}_\xi \mathbf{\Lambda}_x var(\xi)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi'}\gamma$	$\mathbf{\Lambda}_y' \mathbf{A}_\eta' \gamma$	$\frac{\mathbf{A}_\xi \mathbf{\Lambda}_x var(\xi) \mathbf{\Lambda}_y' \mathbf{A}_\eta'}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi'}\gamma$
Regression FSR	γ	γ	$\mathbf{A}_\eta^R \mathbf{\Lambda}_y \gamma$	$\mathbf{A}_\eta^R \mathbf{\Lambda}_y \gamma$
Bartlett FSR	γ	$\frac{var(\xi)}{\mathbf{A}_\xi^B \mathbf{\Sigma}_x \mathbf{A}_\xi^{B\prime}}\gamma$	γ	$\frac{var(\xi)}{\mathbf{A}_\xi^B \mathbf{\Sigma}_x \mathbf{A}_\xi^{B\prime}}\gamma$
Bias avoiding method	γ	γ	γ	γ
Bias correcting method	γ	γ	γ	γ

Note: γ is the true population regression parameter, β is the regression parameter that is obtained when using factor score regression. When $\beta = \gamma$, there is no bias. If it is not equal, there is bias.

2.3.2 Bartlett FSR method

The Bartlett FSR method uses the Bartlett predictor (Bartlett, 1937; Thomson, 1938) to compute the factor scores and then uses these factor scores in a linear regression. The factor scoring matrices $\mathbf{A}_{\xi}^{B}$ and $\mathbf{A}_{\xi}^{B}$ are calculated as follows:

$$\mathbf{A}_{\xi}^{B} = (\mathbf{\Lambda}_{x}^{'}\mathbf{\Theta}_{\delta}^{-1}\mathbf{\Lambda}_{x})^{-1}\mathbf{\Lambda}_{x}^{'}\mathbf{\Theta}_{\delta}^{-1} \quad (2.12)$$

$$\mathbf{A}_{\eta}^{B} = (\mathbf{\Lambda}_{y}^{'}\mathbf{\Theta}_{\epsilon}^{-1}\mathbf{\Lambda}_{y})^{-1}\mathbf{\Lambda}_{y}^{'}\mathbf{\Theta}_{\epsilon}^{-1}, \quad (2.13)$$

with $\mathbf{\Theta}_{\delta}$ and $\mathbf{\Theta}_{\epsilon}$ the covariance matrices of respectively, δ and ϵ. The Bartlett predictor is less known than the regression predictor, but has the advantage that

$$\mathbf{A}_{\xi}^{B}\mathbf{\Lambda}_{x} = (\mathbf{\Lambda}_{x}^{'}\mathbf{\Theta}_{\delta}^{-1}\mathbf{\Lambda}_{x})^{-1}\mathbf{\Lambda}_{x}^{'}\mathbf{\Theta}_{\delta}^{-1}\mathbf{\Lambda}_{x} = 1.$$

This implies that the formulas for the covariances can be simplified. The formula for the variance of ξ stays the same. Combined, this gives the regression coefficients as in Appendix C and Table 2.1. When using Bartlett FSR, there is no longer bias when factor sores are used for the dependent variable. However, now there is bias when factor scores are used for the independent variable. There is also still bias when factor scores are used for both variables.
Again, these results only apply for the unstandardized parameterization. Since the standard deviations are also biased using this method, the standardized regression coefficient will be biased if factor scores are used for any of the variables.

2.3.3 Bias avoiding method

The bias avoiding method was developed by Skrondal and Laake (2001). Based on the results discussed in the previous section, they concluded that one should simply use the regression predictor to predict the factor scores of the independent variable, while one should use the Bartlett predictor to predict the factor scores of the dependent variable. It has already been proven that this works

for the unstandardized parameterization when factor scores are used for only one of the variables, but Skrondal and Laake (2001) showed that this also works when factor scores are used for both variables. In Appendix D, it is proven that this method results in unbiased estimates for all settings, but it has some drawbacks. First of all, one has to determine in advance if a variable will be dependent or independent. Moreover, a variable can only be dependent or independent. Mediational relationships are not possible. Second, this method only works when the unstandardized parameterization is used. When using the standardized parameterization, the regression coefficient estimate will still be biased.

2.3.4 Bias correcting method

The bias correcting method was developed by Croon (2002). In this method, the factor scores are computed using either the regression predictor or the Bartlett predictor. After computing the factor scores, their variances and covariances are calculated. Next, these variances and covariances of the factor scores are used to compute the variances and covariances of the true latent variable scores. Finally, these estimates are used to calculate the regression coefficient. In Appendix E, it is shown how the covariance and variance of the true latent variable scores can be computed. Once these computations have been made, the regression coefficient can be computed as $\beta = \frac{cov(\xi,\eta)}{var(\xi)}$. Since these variances and covariances are unbiased, this results in an unbiased regression coefficient estimate (see Table 2.1). While this process is more complex than the bias avoiding method, it does have some advantages over it. First of all, it works for both the Bartlett and the regression predictor. Secondly, since the variances are no longer biased, the standard deviations necessary to calculate the standardized regression coefficient are also unbiased. This means that the method of Croon (2002) also results in an unbiased regression coefficient when the standardized parameterization is used.

2.4 Standard errors

Skrondal and Laake (2001), Croon (2002) and Lu et al. (2011) only describe how to calculate the regression coefficient, just as in the previous section. To be able to use the methods for hypothesis testing, it is necessary to have a complementary significance test, which requires a standard error and a theoretical distribution. For regression FSR, Bartlett FSR and the bias avoiding method, this is no problem. All three methods perform a regular linear regression after calculating the factor scores. This means that the significance test from the linear regression can be used. This test uses the following standard error:

$$SE = \sqrt{\frac{S^2}{\widehat{var(F_\xi)}(n-1)}},$$

where S^2 is defined as:

$$S^2 = \widehat{var(F_\eta)}(1-r^2)\frac{n-1}{n-(p+1)},$$

with $r = \gamma_z = \frac{sd(\xi)}{sd(\eta)}\gamma$, n is the sample size and p is the number of independent variables. A t-statistic is calculated by dividing the regression coefficient by its standard error. Finally, a p-value is calculated by comparing this t-statistic to a t-distribution with n-(p+1) degrees of freedom. Note, that when performing hypothesis tests, we are dealing with finite samples. For this reason, all population parameters are replaced with their corresponding sample estimates.

For the bias correcting method, the standard error is not so easy to calculate. If one would use the above standard error, then this would be the standard error that coincides with the original, uncorrected regression coefficient. In this case, using the corrected regression coefficient to calculate the t-value would result in an incorrect t-value (and p-value). On the other hand, using the uncorrected regression coefficient would just result in a significance test

for the uncorrected regression coefficient and is again not adequate. Another alternative would be to use the above formula for the standard error, but replace all variances and standard deviations with their corrected versions (Croon, 2002). Unfortunately, this approach implicitly assumes that the true latent scores are directly observable, resulting in an underestimation of the standard error. This suggests that the standard error consists of multiple parts, namely error resulting from the regression itself and error resulting from the factor scores. One way to calculate the error resulting from the factor scores, is to first calculate the prediction error in the factor scores. The prediction error in F_ξ and F_η will be denoted as var_{ε_x} and var_{ε_y} respectively. These prediction errors can be calculated for both the regression and Bartlett predictor. Here, we will only discuss the regression predictor. Skrondal and Rabe-Hesketh (2004, eq. 7.7) showed that the prediction errors can be calculated as follows when using the regression predictor:

$$\begin{aligned} var_{\varepsilon_x} &= \Phi - \Phi' \Lambda' \Sigma_x^{-1} \Lambda \Phi \qquad (2.14) \\ &= var(\xi) - var(\eta). \qquad (2.15) \end{aligned}$$

The last equation is derived from Equation 2.31 in Appendix B. The same derivations can be made for the dependent variable η:

$$var_{\varepsilon_y} = var(\eta) - var(F_\eta). \qquad (2.16)$$

The prediction error in the factor scores is thus the difference between the observed and the corrected variance. Since this is again a variance, the formula from the regular linear regression can be used to calculate the corresponding S^2. From this, it can be derived that the total S^2 consists of three parts:

1. $S^2_{reg} = var(\eta)(1 - r^2)\frac{n-1}{n-(p+1)}$
2. $S^2_y = var_{\varepsilon_y}(1 - r^2)\frac{n-1}{n-(p+1)}$
3. $S^2_x = var_{\varepsilon_x} r^2 \frac{n-1}{n-(p+1)}$.

Note that for S_x^2, r^2 is used instead of $1 - r^2$. This is because the prediction error in the independent variable has more influence on the standard error, as the relation between the variables increases. When the independent variable has no influence on the dependent variable, it also has no influence on the standard error. Now, the total S^2 can be calculated by summing up the three parts:

$$S_{total}^2 = S_{reg}^2 + S_y^2 + S_x^2. \tag{2.17}$$

Finally, a new adjusted standard error can be calculated as:

$$SE = \sqrt{\frac{S_{total}^2}{var(F_\xi)(n-1)}}. \tag{2.18}$$

Using this approximate standard error and the corrected regression coefficient, an approximate t-statistic can be obtained and compared to the theoretical t-distribution with n-(p+1) degrees of freedom. Using this newly developed standard error, the bias correcting method can now be used to perform hypothesis tests about the regression coefficient γ.

2.5 Simulation studies

Two simulation studies are conducted to examine the finite sample performance of these methods and significance tests. In a first study, the methods are studied using item responses that are normally distributed. In the second study, non-normal item responses are used. The results of these studies can be used as guidelines to determine which method to use, depending on the data and research questions.

First, an outline of how the data was simulated is given, followed by a description of the analyses performed on the simulated data and the results of both studies.

2.5.1 Data simulation

Study 1

Before the simulation of the data, a ground truth or population model is defined. The structural model consists of a multivariate regression between three latent independent variables, ξ_1, ξ_2 and ξ_3, and two latent dependent variables η_1 and η_2, resulting in the structural equation:

$$\eta_1 = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_3 + \zeta_1, \tag{2.19}$$

$$\eta_2 = \gamma_3\xi_1 + \gamma_1\xi_2 + \gamma_2\xi_3 + \zeta_2. \tag{2.20}$$

The model is depicted in Figure 2.2.

The simulation of the data consists of two steps, which are carried out using R (R Development Core Team, 2016). In the first step, the true latent variable values of ξ_1, ξ_2, ξ_3, η_1 and η_2 are generated. The variance of the independent variables ξ_1, ξ_2 and ξ_3 is set at 100, while the residual variance of the dependent variables η_1 and η_2 is set at 400. The covariances between all latent variables are 0. To generate data that comply with these parameters, the true latent scores of ξ_1, ξ_2 and ξ_3 are first generated, followed by the regression residuals ζ_1 and ζ_2. Finally, using the structural equations $\eta_1 = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_3 + \zeta_1$ and $\eta_2 = \gamma_3\xi_1 + \gamma_1\xi_2 + \gamma_2\xi_3 + \zeta_2$, the true latent scores on η_1 and η_2 are generated.

In the second step, data is generated for each observed item response x_{ij} and y_{ij}, using the true latent scores and the measurement models, with i referring to the latent variables and j to the items. The measurement models of the latent variables are $y_{ij} = \lambda_{y_{ij}}\eta_i + \epsilon_{ij}$ and $x_{ij} = \lambda_{x_{ij}}\xi_i + \delta_{ij}$. All $\lambda_{x_{ij}}$ and $\lambda_{y_{ij}}$ are set at 1. The residual variances $\Theta_{\epsilon_{ij}}$ of all y_{ij} are set at $\Theta_{\epsilon_{ij}} = \frac{var(\eta_i)(1-CD_{y_i})}{CD_{y_i}}$ and the residual variances $\Theta_{\delta_{ij}}$ of all x_{ij} are set at $\Theta_{\delta_{ij}} = \frac{var(\xi_i)(1-CD_{x_i})}{CD_{x_i}}$, with CD_{y_i} and CD_{x_i} the respective coefficients of determination for the measurement models. All CD_{x_i} and CD_{y_i} are equal and will thus be referred to as CD. To create item responses that are

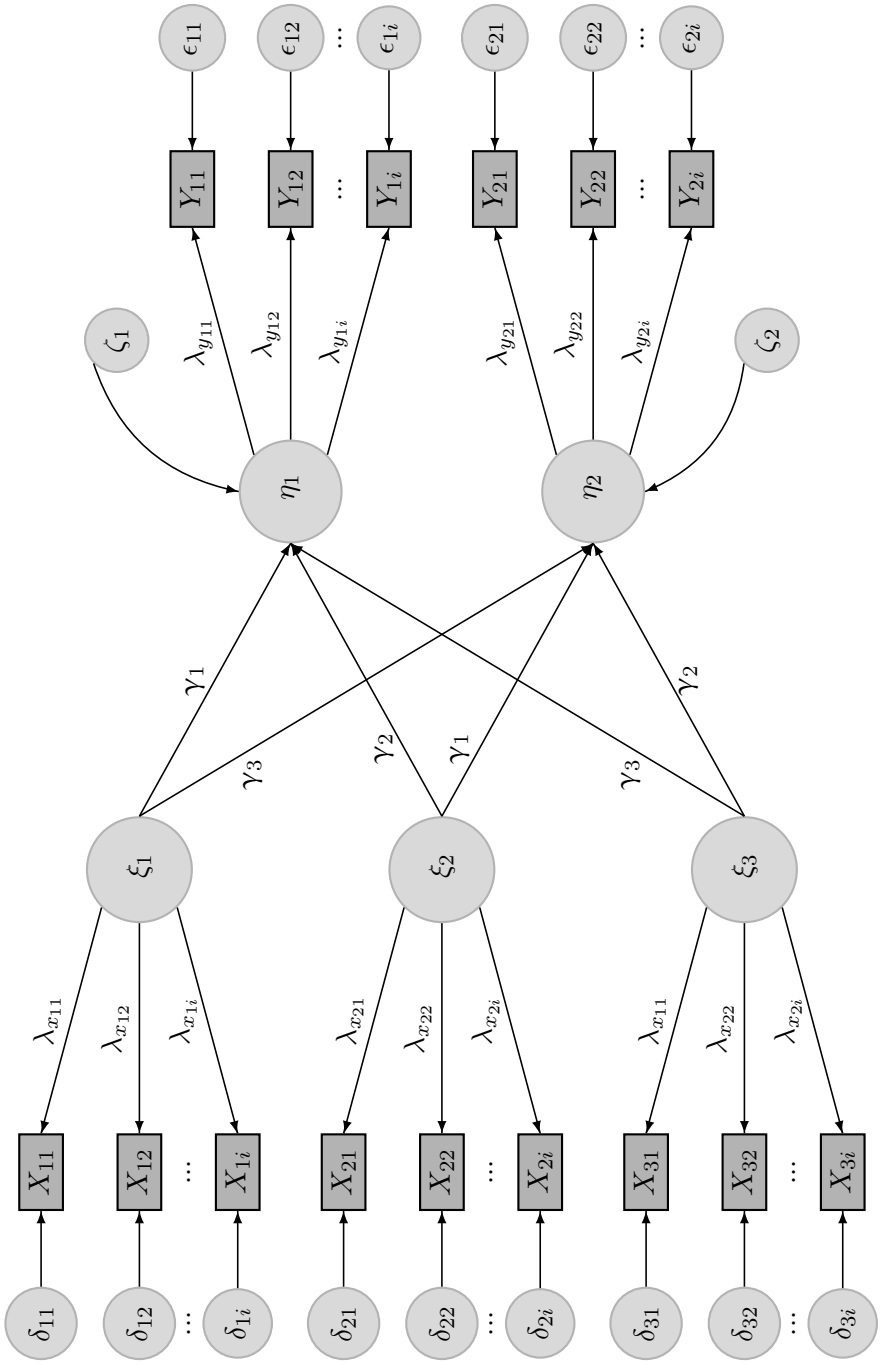


Figure 2.2 The population model

Table 2.2 Summary of the model parameters used in simulation

Model parameter	value
Regression coefficients γ_1	0
γ_2	1.5
γ_3	0.41
coefficient of determination CD	0.3, 0.6 0.7, 0.9 0.95, 0.99
The sample size n	300, 500, 800, 2000
The number of items I	3, 5, 10

normally distributed (skewness = 0, kurtosis = 3), ϵ_{ij} and δ_{ij} are generated from a univariate normal distribution.
The coefficients of determination CD, the regression coefficient γ_i, the sample size n and the number of items I for the two latent variables are varied to create 216 experimental conditions. The sample size is varied to be able to determine the consistency of the methods, while the coefficient of determination is varied to account for the degree of factor score indeterminacy. A higher CD implies a lower degree of factor score indeterminacy (Acito & Anderson, 1986). The regression coefficient is varied to be able to determine the power and type I-error rate. The values used for these parameters can be found in Table 2.2.

Study 2

In the second simulation study, the aim is to compare the methods when the observed item responses are not normally distributed. This means the simulation of the data is done in the same way as in study 1, except for the generation of the ϵ_{ij} and δ_{ij}. To create non-normal item responses, the generation of ϵ_{ij} and δ_{ij} is done in two different ways. To create item responses that are symmetrical, but

have a large kurtosis, ϵ_{ij} and δ_{ij} are generated from a t-distribution with 3 degrees of freedom, multiplied by $\sqrt{\Theta_{\epsilon_{ij}}}$ or $\sqrt{\Theta_{\delta_{ij}}}$, respectively. On average, this results in data with an almost-zero skewness (=0.001) and a kurtosis of about 10.057. To create item responses that are skewed and have a non-normal kurtosis, ϵ_{ij} and δ_{ij} are generated from a χ^2-distributions with 1 degree of freedom, multiplied by $\sqrt{\Theta_{\epsilon_{ij}}}$ or $\sqrt{\Theta_{\delta_{ij}}}$, respectively. On average, this results in data with a skewness of 0.693 and a kurtosis of 4.785. Note that the way the data are simulated implies that the degree of non-normality reduces as the coefficient of determination increases. This is due to the fact that ϵ_{ij} and δ_{ij} have less influence when the factor loadings become stronger.

2.5.2 Analyses

The analysis performed on the data is the same in both studies. Five methods of analyses are performed on the simulated data, namely regression FSR, Bartlett FSR, the bias avoiding method (Skrondal & Laake, 2001), the bias correcting method (Croon, 2002) and SEM. For the SEM analysis, a correctly specified SEM-model is constructed, meaning that it corresponds with the population model used to generate the data. This model is estimated using the simulated dataset, with a "maximum likelihood" estimator. From the results, the regression coefficient $\hat{\gamma}$, and its standard error SE and p-value $\hat{p}$ are obtained, as well as the standardized regression coefficient $\hat{\gamma}_z$. This is done for a 1000 simulated datasets for every simulation condition. Based on all 1000 replications, several performance criteria are calculated, namely the mean regression coefficient estimations $\bar{\gamma}$ and $\bar{\gamma}_z$, the bias using the unstandardized and the bias using standardized parameterization, the empirical standard deviation, the mean square error, the mean standard error and the standard error bias for the regression coefficient and the power of the statistical test. For the conditions with a regression coefficient of 0, the type I-error rate is calculated instead of the power. For the four FSR methods, a FA is performed for all latent variables.

Hereafter, factor scores are calculated, using the Bartlett and the regression predictor. Next, the regression factor scores are used in two linear regressions for the regression FSR and the Bartlett factor scores are used for the Bartlett FSR. For the bias avoiding method, two linear regressions are performed, using the Bartlett factor scores for the dependent variables and the regression factor scores for the independent variables. Finally, for the bias correcting method, the regression coefficient and standard errors are calculated using the formulas described above. Again, the regression coefficients $\hat{\gamma}$ and $\hat{\gamma}_z$ and the standard error SE and p-value $\hat{p}$ for $\hat{\gamma}$ were retained. This was repeated 1000 times and the same performance criteria were calculated. The criteria are summarized in Table 2.3.

2.5.3 Analysis of the results

Study 1

To compare the five methods with regard to the bias using the unstandardized and standardized parameterization, efficiency, MSE, standard error bias, type I-error rate and power, an ANOVA was performed for each of these performance measures. The independent variables were the design factors, namely the sample size, coefficient of determination, method, number of items and the value of gamma.
All possible two-way interactions were also included in the analyses. This resulted in 15 predictors for each model. The results of this ANOVAs can be found in Table 2.4. Due to space constraints, only the informative effects are discussed in the results section.

Study 2

The same ANOVAs were performed when the data were not normally distributed. However, one extra independent variable was added, namely the degree of non-normality. This resulted in 21 predictors. The results can be found in Table 2.6.

Table 2.3 Summary of the performance criteria, with R the number of successful replications.

Criteria	Formula
$\bar{\gamma}$	$\frac{1}{R}\sum_{i=1}^{R}\hat{\gamma}_i$
$\bar{\gamma}_z$	$\frac{1}{R}\sum_{i=1}^{R}\hat{\gamma}_{z_i}$
Bias	$\frac{1}{R}\sum_{i=1}^{R}(\hat{\gamma}_i-\gamma)$
Relative bias	$\frac{\text{Bias}}{\gamma}$
Empirical standard deviation (ESD)	$\sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(\hat{\gamma}_i-\bar{\gamma})^2}$
Mean square error (MSE)	$\frac{1}{R}\sum_{i=1}^{R}(\hat{\gamma}_i-\gamma)^2$
Mean standard error (MSTE)	$\frac{1}{R}\sum_{i=1}^{R}\text{SE}$
Standard error bias (SEB)	$\frac{1}{R}\sum_{i=1}^{R}(\text{MSTE}-\text{ESD})$
Power/Type I-error	$\frac{1}{R}\sum_{i=1}^{R}(\hat{p}_i<0.05)$

2.5.4 Results Study 1

The results are discussed per statistical performance criterion. The proportion of successful replications, the bias and efficiency results are discussed first, followed by the MSE, standard error bias, type I-error rate and power. In Table 2.5, a comparison between the methods for all performance criteria is given.

Proportion of successful replications

The first performance measure that was considered was the proportion of successful replications for each method. The proportion of

Table 2.4 ANOVA-models, using normally distributed data

Effect	Df	Successful replications	Bias	Standardized bias	Efficiency
I	2	1253.97 ***	43.37 ***	105.00 ***	271.22 ***
γ	2	0.00	264.20 ***	722.37 ***	1192.35 ***
CD	5	839.31 ***	154.10 ***	412.64 ***	3131.66 ***
Method	4	48.24 ***	265.33 ***	230.36 ***	310.78 ***
n	3	1.41	0.16	0.18	15202.77 ***
Method*CD	20	43.35 ***	74.21 ***	67.21 ***	127.58 ***
Method*γ	8	0.00	131.22 ***	113.74 ***	3.92 **
Method*I	8	40.24 ***	17.73 ***	16.52 ***	32.96 ***
Method*n	12	2.22 *	0.02	0.012	12.15 ***
CD*γ	10	0.00	74.66 ***	202.22 ***	389.46 ***
CD*I	10	800.24 ***	10.60 ***	26.21 ***	204.36 ***
CD*n	15	7.21 ***	0.03	0.04	181.63 ***
γ *I	4	0.00	20.55 ***	50.20 ***	0.88
γ *n	6	0.00	0.70	0.33	50.27 ***
I*n	6	3.11 *	0.08	0.08	70.98 ***

Effect	Df	MSE	Standard error bias	Type I-error rate	Power
I	2	38.98 ***	14.13 ***	47.11 ***	55.84 ***
γ	2	144.13 ***	1122.45 ***	/	921.72 ***
CD	5	164.28 ***	531.54 ***	2.92 +	177.54 ***
Method	4	40.55 ***	62.15 ***	9.33 ***	0.22
n	3	65.76 ***	79.20 ***	38.13 ***	429.60 ***
Method*CD	20	25.57 ***	32.43 ***	1.42	0.17
Method*γ	8	44.50 ***	65.25 ***	/	0.22
Method*I	8	8.14 ***	1.64	0.16	0.05
Method*n	12	0.37	1.35	0.22	0.04
CD*γ	10	83.48 ***	355.61 ***	/	177.66 ***
CD*I	10	23.48 ***	12.22 ***	15.40 ***	24.47 ***
CD*n	15	3.33 ***	28.90 ***	3.41 ***	48.33 ***
γ*I	4	18.90 ***	0.97	/	56.01 ***
γ*n	6	1.01	50.96 ***	/	429.17 ***
I*n	6	0.65	0.99	5.89 ***	13.17 ***
Residuals	2044			634	1339

Note:*** < 0.000, **0.000-0.001, *0.001-0.01, +0.01-0.05

Table 2.5 Comparison between the five methods per performance criteria

	regression FSR	Bartlett FSR	Bias avoiding	Bias correcting	SEM
Number of Succesful replications	2	2	2	2	1
Bias	4	5	1	1	1
Standardized bias	3	3	3	1	1
Efficiency	2	1	3	4	4
MSE	4	5	1	1	1
Standard error bias	4	2	5	3	1
Type I-error	1	1	1	1	1
Power	1	1	1	1	1

Note: The numbers indicate the performance of the methods in relation to each other for each performance criteria. A score of 1 means the method performed best on that particular performance critera, while a score of 5 means the methods performance the worst.

successful replications is very high for all methods and conditions, namely 0.989 or higher. The proportion of succesful replications of SEM is even higher than the proportions of the other four methods, which all have the same proportions.

Bias

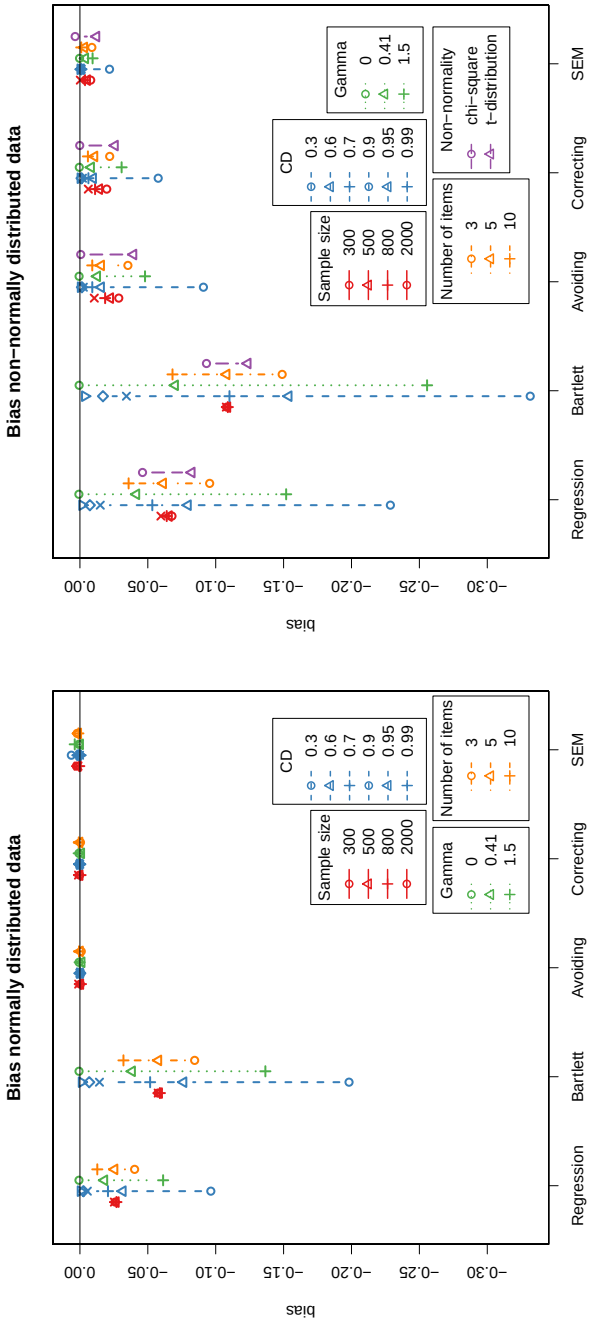
As can be seen in Figure 2.3a, only the regression FSR and the Bartlett FSR method are biased, while the three remaining methods are not. More specifically, both methods underestimate the regression coefficient and the Bartlett FSR method is more severely biased than the regression FSR method. Moreover, the bias is not influenced by the sample size, implying that both the regression and Bartlett FSR methods are also inconsistent. The bias does disappear with an increasing factor loading and when γ is equal to 0. When the number of items goes up, the bias also declines, but has not disappeared completely when the number of items reaches 10. The three other methods, namely the bias avoiding, bias correcting and SEM method, exhibit, as expected, very little bias. The three methods perform very similar.

Bias using standardized parameterization

The patterns change slightly when the standardized parameterization is used (see Figure 2.4a). Now, the bias avoiding method also underestimates the regression coefficient and is inconsistent. There is also no longer a difference between the Bartlett FSR, regression FSR and bias avoiding methods. All other effects can be interpreted in the same way as in the unstandardized parameterization.

Efficiency

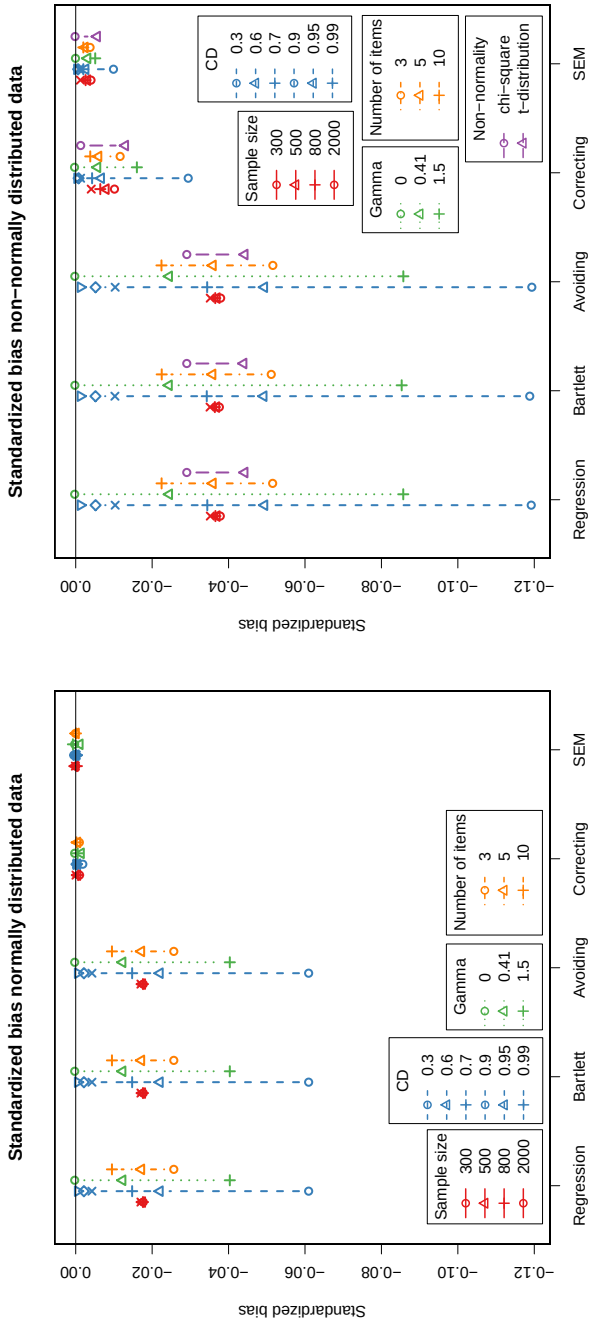
While regression FSR and Bartlett FSR are the most biased methods, they are also the most efficient methods. The Bartlett FSR method is even slightly more efficient than the regression FSR



(a) Normally distributed data.

(b) Non-normally distributed data.

Figure 2.3 The influence of sample size, CD, number of items and the value of γ on the bias using the unstandardized parameterization, in interaction with the method.



(a) Normally distributed data.

(b) Non-normally distributed data.

Figure 2.4 The influence of sample size, CD, number of items and the value of γ on the bias using the standardized parameterization, in interaction with the method.

method. The three other methods have very similar standard errors. However, the bias avoiding method is slightly more efficient than the other two when the coefficient of determination is low. It is also important to note that the differences between the methods disappear as the coefficient of determination increases.

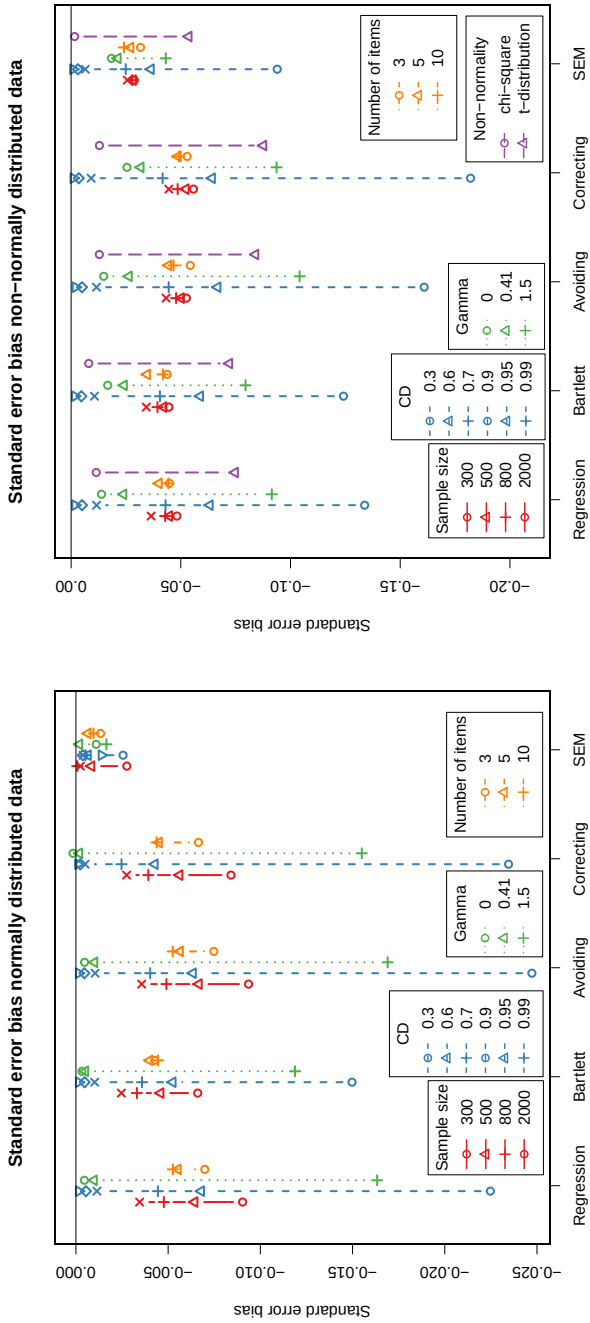
Mean Square Error

The regression and Bartlett FSR methods have a high MSE, as compared to the other three methods, with the Bartlett FSR having the worst MSE. The MSE of the other three methods is very similar to each other. The differences between the five methods disappear when the CD approaches 1 or when the value of γ approaches 0.

Standard error bias

With regard to the standard error bias, there is a large difference between the methods. In Figure 2.5a, it can be seen that all five methods show bias in the standard error. However, the SEM method clearly has the smallest standard error bias. The four alternative methods are more severely biased with regard to the standard error. On average, the Bartlett FSR method gives the second best estimation of the standard error, followed by the bias correcting and regression FSR methods.

Over all methods, a higher sample size, a higher CD or more items, lowers the standard error bias, while a larger value of γ increases the standard error bias. When the CD approaches 1, all differences between the methods disappear and when the value of γ increases, the differences between the methods also increases. It is important to note that when there is no effect ($\gamma = 0$), the standard error bias almost completely disappears for all five methods. This means the standard error bias will have little effect on the type I-error rate.



(a) Normally distributed data.

(b) Non-normally distributed data.

Figure 2.5 The influence of sample size, coefficient of determination, number of items and the value of γ on the standard error bias, in interaction with the method. Note that the scale of the y-axis is not the same in panel (a) and (b).

Type I-error

The type I-error rates for the regression FSR, Bartlett FSR and bias avoiding method are very similar, while SEM has a slightly higher type I-error rate and the bias correcting method has a slightly lower type I-error rate. While these differences are statistically significant, for all five methods the type-I error rates are around the expected value of 0.05 . In practice, this means these differences are not really relevant.

Power

All five methods have approximately the same power. There are only small differences when the coefficient of determination is very weak. In that case, SEM has the lowest power, followed by the bias correcting method. The other methods have the same power.

2.5.5 Results Study 2

The results of the second study are described in relation to the first study. The impact of the kind of non-normality is also discussed.

Proportion of successful replications

When the data are not normally distributed, the proportion of successful replications is a lot lower and the differences between the methods are larger. Especially the proportion of successful replications of SEM is very low as compared to the other four methods. This is mostly due to the conditions where the coefficient of determination is very low (CD=0.3) or the number of items is low (I=3). The proportion of the bias correcting method is the second lowest, followed by the bias avoiding method. The regression FSR method has the highest proportion of successful replications. All these effects are largely due to the conditions where a t-distribution was used to simulate the item responses. The χ^2-distribution seems to have less effect. This could be due to the fact that the kur-

Table 2.6 ANOVA-models, using non-normal data

Effect	Df	Successful replications	Bias	Standardized bias	Efficiency
I	2	124.49 ***	168.59 ***	270.76 ***	240.69 ***
γ	2	0.00	1215.43 ***	2449.80 ***	2974.83 ***
CD	5	1219.83 ***	686.33 ***	1266.31 ***	6770.51 ***
Method	4	183.97 ***	528.00 ***	525.92 ***	190.16 ***
n	3	79.03 ***	4.31 *	4.24 *	2657.27 ***
NN	1	2100.28 ***	299.53 ***	342.32 ***	8044.76 ***
Method*CD	20	121.49 ***	110.43 ***	122.62 ***	15.08 ***
Method*γ	8	0.00	262.68 ***	264.49 ***	3.55 **
Method*I	8	28.19 ***	22.15 ***	25.99 ***	3.15 *
Method*n	12	1.83 +	0.55	0.20	0.49
Method*NN	4	153.36 ***	5.82 **	7.09 ***	3.15 *
CD*γ	10	0.00	340.41 ***	615.95 ***	644.05 ***
CD*I	10	36.24 ***	33.25 ***	45.76 ***	32.18 ***
CD*n	15	36.33 ***	2.64 **	1.77	62.45 ***
CD*NN	5	1087.06 ***	111.44 ***	92.32 ***	1874.78 ***
γ *I	4	0.00	77.43 ***	126.66 ***	4.54 *
γ *n	6	0.00	1.88	1.33	2.57 +
γ*NN	2	0.00	152.23 ***	165.91 ***	1060.31 ***
I*n	6	3.39 *	1.07	0.42	4.04 **
I*NN	2	61.01 ***	22.12 ***	14.55 ***	8.72 **
n*NN	3	54.25 ***	7.35 ***	2.89 +	2.53

Effect	Df	MSE	Standard error bias	Type I-error rate	Power
I	2	96.50 ***	4.57 +	3.83 +	321.30 ***
γ	2	824.44 ***	2473.21 ***	/	3549.41 ***
CD	5	935.59 ***	2733.13 ***	378.18 ***	1307.85 ***
Method	4	48.17 ***	95.91 ***	35.50 ***	0.72
n	3	70.00 ***	4.67 *	31.634 ***	810.00 ***
NN	1	657.45 ***	6300.45 ***	1804.36 ***	365.06 ***
Method*CD	20	20.40 ***	35.43 ***	31.72 ***	1.49
Method*γ	8	64.50 ***	70.14 ***	/	0.99
Method*I	8	3.88 **	2.19 +	1.27	0.57
Method*n	12	0.48	0.68	1.21	0.05
Method*NN	4	5.28 **	22.85 ***	33.79 ***	2.11
CD*γ	10	408.33 ***	524.64 ***	/	926.40 ***
CD*I	10	38.39 ***	0.74	15.70 ***	111.62 ***
CD*n	15	16.93 ***	1.48	17.76 ***	114.89 ***
CD*NN	5	308.80 ***	1449.50 ***	340.35 ***	126.93
γ *I	4	55.55 ***	23.48 ***	/	198.07 ***
γ *n	6	4.79	9.81 ***	/	662.82 ***
γ*NN	2	265.38 ***	1037.45 ***	/	158.74 ***
I*n	6	1.13	0.64	1.36	22.92 ***
I*NN	2	3.88 **	8.01 **	0.31	26.26 ***
n*NN	3	11.56 ***	0.71	74.15 ***	13.08
Residuals	4187			1339	2763

Note:*** < 0.000, **0.000-0.001, *0.001-0.01, +0.01-0.05, NN = Non-normality

tosis is larger when using the t-distribution than when using the χ^2-distribution.

Bias

With regard to the bias, there are two main shifts in the patterns (Figure 2.3b). Firstly, now, there is a difference between SEM, the bias avoiding and bias correcting method. SEM still shows almost no bias, while the bias correcting and bias avoiding method do show a little bias. The bias avoiding method is more biased than the bias correcting method. The regression FSR and Bartlett FSR still show a large amount of bias. Secondly, the sample size does have an influence for SEM, the bias avoiding and bias correcting method. Because of the larger bias in these methods, it can now be seen that these methods are consistent. Again, these patterns are largely caused by the conditions using the t-distribution.

Bias using standardized parameterization

The patterns with regard to the bias using standardized parameterization change in two ways (see Figure 2.4b). Firstly, both SEM and the bias correcting method show a small bias in the standardized regression parameter, with SEM having the smallest bias. Secondly, the sample size does have an influence on the bias of the bias correcting and SEM method when the data is not normally distributed. When the sample size increases, the bias decreases.

Efficiency

The efficiency is a lot lower when the data are not normal, especially when the t-distribution was used to simulate the data.

Mean Square Error

While the MSE of SEM, bias correcting and bias avoiding method were very similar to each other when the data was normally distributed, now there are small differences. SEM has the smallest

MSE, closely followed by the bias correcting method. This is again due to the high kurtosis of the conditions simulated with the t-distribution. All other effects remain the same.

Standard error bias

When the data are not normally distributed, the standard error bias of the bias correcting method is the worst in some conditions (see Figure 2.5b), especially when the CD or the value of γ is very low.

Type I-error

When the data are not normally distributed, especially when there is a high kurtosis, the type I-error rate is larger than the expected value of 0.05 for all methods. Especially the bias correcting method seems to have a large type I-error rate, when the coefficient of determination is very low, namely 0.30.

Power

As can be expected, the power seems to be lower when the data are not normally distributed. Table 2.6 shows that the method does not have a significant influence on the power, when the data are not normally distributed.

2.6 Discussion

In this paper, an overview was given of four methods for factor score regression, namely regression FSR, Bartlett FSR, the bias avoiding method (Skrondal & Laake, 2001) and the bias correcting method (Croon, 2002). The four methods were described and their statistical properties were discussed on the population level. Since there was no adequate standard error available for the bias correcting method, a new standard error was developed. To be able to determine the statistical properties of the methods in finite samples and

to evaluate the performance of the newly developed standard error, two Monte Carlo simulation studies were performed.
The simulation studies showed that the regression FSR method does not perform well. This confirms the general expectations found in the literature (e.g. Bollen, 1989; Croon, 2002; Lastovicka & Thamodaran, 1991; Lewis, 2005; Shevlin et al., 1997; Skrondal & Laake, 2001). It also complies with the results of Lu et al. (2011). The method is biased for both the standardized and unstandardized parameterization and is inconsistent. The method does have a high efficiency, but it also has a high MSE. This means that the high efficiency cannot compensate for the high bias. Moreover, the estimates of the standard error have the second highest bias observed.

The Bartlett FSR method performs even worse than the regression FSR method, with a comparable bias using the standardized parameterization, but a higher unstandardized bias and MSE. It does have a lower standard error bias. It is also the most efficient method, but at the same time it has the highest bias and MSE of all methods.

The first corrected method, namely the bias avoiding method of Skrondal and Laake (2001), only performs slightly better than the regression FSR and Bartlett FSR methods. It is unbiased when using the unstandardized parameterization, but it is still biased when using the standardized parameterization. This result highlights the fact that standardized and unstandardized regression coefficients do not always behave in the same way (Kim & Mueller, 1976). The standard error bias is the largest of all methods and it has the same power and type I-error rate as regression FSR and Bartlett FSR. It can be concluded that this method only outperforms regression FSR and Bartlett FSR with regard to the unstandardized bias and is outperformed by SEM and the bias correcting method.

The second corrected method, the bias correcting method, performs better than the bias avoiding method. It is unbiased for both the standardized and unstandardized parameterization and

has the highest power. When the observed item responses are not normally distributed, it does show a slightly larger bias than the SEM method, but its proportion of successful replications is also much higher.

With regard to the standard error bias, only SEM and Bartlett FSR do better. This result shows that the newly developed standard error is reliable and even performs better than the regular standard error used in the bias avoiding method. However, it is important to note that when the data are not normally distributed and the factor loadings are very weak (CD=0.3), the standard error bias goes up, resulting in a higher type I-error rate. In all other conditions, the estimate of the standard error is reliable. The method does have the lowest efficiency, but also has the lowest MSE, meaning that the low efficiency does not have much influence.

The SEM method performs very similar to the bias correcting method. When the data are normally distributed, SEM has the same bias when using the unstandardized and standardized parameterization, efficiency, MSE, power and type I-error rate. When the data are not normally distributed, SEM has a lower bias and MSE, but also has a very low proportion of successful replications. This means that, although SEM gives less biased regression coefficients, the chance that the model will not converge, is also a much larger for SEM. On the other hand, there is almost no standard error bias when using the SEM method.

Overall, it can be concluded that only the bias correcting method is a suitable alternative for SEM. The method performs similar to the SEM method with regard to bias, efficiency, MSE, power and type I-error rate. It does have more standard error bias than SEM, but it has the second lowest standard error bias of the four FSR methods. It also has more successful replications than SEM when the data are not normally distributed. The method does have some drawbacks in comparison to SEM, because of its two-step nature. First, at this moment, there are no overall fit indices available. Second, the method cannot be used for all possible structural models. For

example, the method cannot handle mediational relationships and is not applicable to non-recursive methods. However, it is the intention of the authors to extend the method to be applicable to the full SEM-model and to develop a set of fit indices for this method.

A second conclusion that can be made is that factor indeterminacy plays a great role in deciding which method to use. The simulation study showed that if the factor indeterminacy is low (i.e. a coefficient of determination of 0.99) the differences between the methods disappear completely, on all performance criteria. As a result, there is no longer a problem with performing a conventional factor score regression. This implies that it is important to first determine the factor score indeterminacies by use of indeterminacy indices. If these indices suggest that the indeterminacies are very low, any of the methods can be used. If these indices suggest the indeterminacies are moderate or high, one should use SEM or the bias correcting method.

In this paper it was shown that the bias correcting method of Croon (2002) is a reliable and unbiased method to perform a factor score regression. A new and reliable standard error was also developed, meaning that the bias correcting method can now be used by applied users to perform significance tests. However, performing the method is a rather complex and technical process. For this reason, software to perform the method will be developed and made available for the applied users in the near future.

References

Acito, F., & Anderson, R. D. (1986). A simulation study of factor score indeterminacy. *Journal of Marketing Research*, *23*(2), 111–118. doi: 10.2307/3151658

Bartlett, M. (1937, jul). The statistical conception of mental factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*, 78–117. doi: 10.1177/0049124187016001004

Bollen, K. (1989). *Structural equations with latent variabels.* New York: NY: Wiley.

Bollen, K., & Hoyle, R. H. (2012). Latent variables in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 59–67). the Guilford Press.

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. doi: 10.1037//1082-989X.6.4.430

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social science.s* (pp. 85–112). New York: Seminar Press. doi: 10.1002/j.2333-8504.1970.tb00783.x

Kim, J.-O., & Mueller, C. W. (1976). Standardized and unstandardized coefficients in causal analysis: An expository note. *Sociological Methods & Research*, *4*(4), 423–438. doi: 10.1177/004912417600400402

Lastovicka, J. L., & Thamodaran, K. (1991). Common factor score estimates in multiple regression problems. *Journal of Marketing Research*, *28*(1), 105–112. doi: 10.2307/3172730

Lewis, J. B. (2005, jul). Estimating regression models in which the dependent variable Is based on estimates. *Political Analysis*, *13*(4), 345–364. doi: 10.1093/pan/mpi026

Lu, I. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011, sep). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, *28*(3),

258–268. doi: 10.1016/j.ijresmar.2011.03.006

Lu, I. R., & Thomas, D. R. (2008, jul). Avoiding and correcting bias in score-based latent variable regression with discrete manifest items. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 462–490. doi: 10.1080/10705510802154323

Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models : A comparison with regression based on IRT scores. *Structural equation modeling: A multidisciplinary Journal*, *12:2*(January 2014), 263–277. doi: 10.1207/s15328007sem1202

Maraun, M. D. (1996). Metaphor taken as math: Indeterminancy in the factor analysis model. *Multivariate Behavioral Research*, *31*(4), 517–538. doi: 10.1207/s15327906mbr3104_6

Mulaik, S. A. (1972). Factor scores and factor indeterminacy. In S. A. Mulaik (Ed.), *Foundations of factor analysis.* New York: McGraw-Hill Book Company.

R Development Core Team. (2016). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical computing.

Shevlin, M., Miles, J. N. V., & Bunting, B. P. (1997). Summated rating scales. A Monte Carlo investigation of the effects of reliability and collinearity in regression models. *Personality and Individual Differences*, *23*(4), 665–676. doi: 10.1016/S0191-8869(97)00088-3

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. doi: 10.1007/BF02296196

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, *44*(2), 157–167. doi: 10.1007/BF02293967

Thomson, G. (1934). The meaning of i in the estimate of g. *British Journal of Psychology*, *25*, 92–99. doi: 10.1111/j.2044-8295.1934.tb00728.x

Thomson, G. (1938, feb). Methods of Estimating Mental Factors. *Nature*, *141*(3562), 246–246. doi: 10.1038/141246a0

Thurstone, L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press. doi: 10.1037/h0075959

Appendices

2.A Regression coefficient in the general case

In this appendix, we derive the relationship between β and γ in the general case. We make a distinction between three scenarios:
1) both the independent and dependent variable are latent variables
2) only the independent variable is a latent variable
3) only the dependent variable is a latent variable"

2.A.1 Independent and dependent latent variable

When performing the linear regression with factor scores, the regression coefficient becomes $\beta = \frac{cov(F_\xi, F_\eta)}{var(F_\xi)}$, which is not necessarily the same as the true regression coefficient. The relationship between γ and β can best be understood if we work out the covariance and variances of the factor scores. The covariance can be written as :

$$\begin{aligned} cov(F_\xi, F_\eta) &= cov(\mathbf{A}_\xi \mathbf{x}, \mathbf{A}_\eta \mathbf{y}) \\ &= \mathbf{A}_\xi cov(\mathbf{x}, \mathbf{y}) \mathbf{A}_\eta^{'} \\ &= \mathbf{A}_\xi cov(\mathbf{\Lambda}_x \xi + \boldsymbol{\delta}, \mathbf{\Lambda}_y \eta + \boldsymbol{\epsilon}) \mathbf{A}_\eta^{'} \\ &= \mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi + \boldsymbol{\delta}, \eta + \boldsymbol{\epsilon}) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'} \\ &= \mathbf{A}_\xi \mathbf{\Lambda}_x [cov(\xi, \eta) + cov(\xi, \boldsymbol{\epsilon}) + \\ &\quad cov(\eta, \boldsymbol{\epsilon}) + cov(\boldsymbol{\delta}, \boldsymbol{\epsilon})] \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'} \\ &= \mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}. \end{aligned} \tag{2.21}$$

The variance can be written as:

$$\begin{aligned} var(F_\xi) &= var(\mathbf{A}_\xi \mathbf{x}) \\ &= \mathbf{A}_\xi var(\mathbf{x}) \mathbf{A}_\xi^{'} \\ &= \mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}, \end{aligned} \tag{2.22}$$

where $\mathbf{\Sigma}_x$ is the variance of $\mathbf{x}$. Based on these calculations, the regression coefficient becomes:

$$\beta = \frac{cov(F_\xi, F_\eta)}{var(F_\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x var(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} \gamma. \tag{2.23}$$

2.A.2 Independent latent variable

When factor sores are only used for the independent variable, the regression coefficient becomes $\beta = \frac{cov(F_\xi, \eta)}{var(F_\xi)}$. Again, we work out the covariance and variances of the factor scores:

$$\begin{aligned} var(F_\xi) &= \mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'} && (2.24) \\ cov(F_\xi, \eta) &= cov(\mathbf{A}_\xi \mathbf{x}, \eta) \\ &= \mathbf{A}_\xi cov(\mathbf{x}, \eta) \\ &= \mathbf{A}_\xi cov(\mathbf{\Lambda}_x \xi, \eta) \\ &= \mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta) && (2.25) \end{aligned}$$

Based on these calculations, the estimated regression coefficient becomes:

$$\beta = \frac{cov(F_\xi, \eta)}{var(F_\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x var(\xi, \eta)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} \gamma. \tag{2.26}$$

2.A.3 Dependent latent variable

When only the dependent variable consists of factor scores, the regression coefficient becomes $\beta = \frac{cov(\xi, F_\eta)}{var(\xi)}$. Again, we work out the covariance and variances of the factor scores:

$$\begin{aligned} cov(\xi, F_\eta) &= cov(\xi, \mathbf{A}_\eta \mathbf{y}) \\ &= cov(\xi, \mathbf{y}) \mathbf{A}_\eta \\ &= cov(\xi, \mathbf{\Lambda}_y \eta) \mathbf{A}_\eta^{'} \\ &= cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'} && (2.27) \end{aligned}$$

Based on these calculations, the estimated regression coefficient becomes:

$$\beta = \frac{cov(\xi, F_\eta)}{var(\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta)}{var(\xi)} = \mathbf{A}_\xi \mathbf{\Lambda}_x \gamma. \tag{2.28}$$

2.B Regression coefficient using regression FSR

In this appendix, we derive the relationship between β and γ when we use the regression predictor to calculate the factor scores. When using the regression predictor, the factor scoring matrices $\mathbf{A}_\xi^R$ and $\mathbf{A}_\eta^R$ are computed as follows:

$$\mathbf{A}_\xi^R = \mathbf{\Phi}\mathbf{\Lambda}_x^{'}\mathbf{\Sigma}_x^{-1} = var(\xi)\mathbf{\Lambda}_x^{'}\mathbf{\Sigma}_x^{-1}, \tag{2.29}$$

$$\mathbf{A}_\eta^R = var(\eta)\mathbf{\Lambda}_y^{'}\mathbf{\Sigma}_y^{-1}. \tag{2.30}$$

This means the formula for the variance of F_ξ can be simplified:

$$\begin{aligned} var(F_\xi^R) &= \mathbf{A}_\xi^R \mathbf{\Sigma}_x \mathbf{A}_\xi^{R'} \\ &= var(\xi)\mathbf{\Lambda}_x^{'}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_x\mathbf{A}_\xi^{R'} \\ &= var(\xi)\mathbf{\Lambda}_x^{'}\boldsymbol{I}\mathbf{A}_\xi^{R'} \\ &= var(\xi)\mathbf{\Lambda}_x^{'}\mathbf{A}_\xi^{R'} = \Phi^{'}\Lambda^{'}\Sigma_x^{-1}\Lambda\Phi \\ &= \mathbf{A}_\xi^R \mathbf{\Lambda}_x var(\xi). \end{aligned} \tag{2.31}$$

The formulas for the covariances stay unchanged. Based on these calculations, the regression coefficient can be recalculated. When both variables are latent, the regression coefficient becomes:

$$\begin{aligned} \beta &= \frac{cov(F_\xi, F_\eta)}{var(F_\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta)\mathbf{\Lambda}_y^{'}\mathbf{A}_\eta^{'}}{\mathbf{A}_\xi^R \mathbf{\Lambda}_x var(\xi)} \\ &= \frac{cov(\xi, \eta)\mathbf{\Lambda}_y^{'}\mathbf{A}_\eta^{'}}{var(\xi)} = \mathbf{\Lambda}_y^{'}\mathbf{A}_\eta^{'}\gamma. \end{aligned} \tag{2.32}$$

When only the independent variable is latent, the regression coeffi-

cient becomes:

$$\beta = \frac{cov(F_\xi, \eta)}{var(F_\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta)}{\mathbf{A}_\xi^R \mathbf{\Lambda}_x var(\xi)} = \frac{cov(\xi, \eta)}{var(\xi)} = \gamma. \tag{2.33}$$

When only the dependent variable is latent, the regression coefficient becomes:

$$\beta = \frac{cov(\xi, F_\eta)}{var(\xi)} = \frac{\mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta)}{var(\xi)} = \mathbf{A}_\xi \mathbf{\Lambda}_x \gamma. \tag{2.34}$$

2.C Regression coefficient using Bartlett FSR

In this appendix, we derive the relationship between β and γ when we use the Bartlett predictor to calculate the factor scores. The Bartlett predictor has the advantage that $\mathbf{A}_\xi^B \mathbf{\Lambda}_x = 1$ and $\mathbf{A}_\eta^B \mathbf{\Lambda}_y = 1$. This implies that the formulas for the covariances can be simplified:

$$\begin{aligned} cov(F_\xi, F_\eta) &= \mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'} = cov(\xi, \eta) \\ cov(F_\xi, \eta) &= \mathbf{A}_\xi \mathbf{\Lambda}_x cov(\xi, \eta) = cov(\xi, \eta) \\ cov(\xi, F_\eta) &= cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'} = cov(\xi, \eta) \end{aligned} \tag{2.35}$$

The formula for the variance of ξ stays the same. Combined, this gives the following regression coefficient, when both variables are latent:

$$\beta = \frac{cov(F_\xi, F_\eta)}{var(F_\xi)} = \frac{cov(\xi, \eta)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} = \frac{var(\xi)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} \gamma. \tag{2.36}$$

When only the independent variable is latent, the regression coefficient becomes:

$$\beta = \frac{cov(F_\xi, \eta)}{var(F_\xi)} = \frac{cov(\xi, \eta)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} = \frac{var(\xi)}{\mathbf{A}_\xi \mathbf{\Sigma}_x \mathbf{A}_\xi^{'}} \gamma. \tag{2.37}$$

When only the dependent variable is latent, the regression coefficient becomes:

$$\beta = \frac{cov(\xi, F_\eta)}{var(\xi)} = \frac{cov(\xi, \eta)}{var(\xi)} = \gamma. \tag{2.38}$$

2.D Regression coefficient using the bias avoiding method

In this appendix, we derive the relationship between β and γ when we use the bias avoiding method, which uses the regression predictor to predict the factor scores of the independent variable and the Bartlett predictor to predict the factor scores of the dependent variable. The covariances between the two variables then becomes:

$$\begin{aligned}
cov(F_\xi^R, F_\eta^B) &= cov(\mathbf{A}_\xi^R \mathbf{x}, \mathbf{A}_\eta^B \mathbf{y}) \\
&= \mathbf{A}_\xi^R cov(\mathbf{x}, \mathbf{y}) \mathbf{A}_\eta^{B'} \\
&= \mathbf{A}_\xi^R cov(\mathbf{\Lambda}_x \xi + \boldsymbol{\delta}, \mathbf{\Lambda}_y \eta + \boldsymbol{\epsilon}) \mathbf{A}_\eta^{B'} \\
&= \mathbf{A}_\xi^R \mathbf{\Lambda}_x cov(\xi + \boldsymbol{\delta}, \eta + \boldsymbol{\epsilon}) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{B'} \\
&= \mathbf{A}_\xi^R \mathbf{\Lambda}_x [cov(\xi, \eta) + cov(\xi, \boldsymbol{\epsilon}) \\
&\quad + cov(\eta, \boldsymbol{\epsilon}) + cov(\boldsymbol{\delta}, \boldsymbol{\epsilon})] \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{B'} \\
&= \mathbf{A}_\xi^R \mathbf{\Lambda}_x cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{B'} \\
&= \mathbf{A}_\xi^R \mathbf{\Lambda}_x cov(\xi, \eta) \qquad (2.39) \\
cov(F_\xi^R, \eta) &= \mathbf{A}_\xi^R \mathbf{\Lambda}_x cov(\xi, \eta) \qquad (2.40) \\
cov(\xi, F_\eta^B) &= cov(\xi, \eta) \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{B'} = cov(\xi, \eta), \qquad (2.41)
\end{aligned}$$

while the variance can be written as:

$$var(F_\xi^R) = \mathbf{A}_\xi^R \mathbf{\Lambda}_x var(\xi). \tag{2.42}$$

Combined, this gives the following regression coefficient, when both variables are latent:

$$\beta = \frac{cov(F_\xi^R, F_\eta^B)}{var(F_\xi^R)} = \frac{\mathbf{A}_\xi^R \mathbf{\Lambda}_x cov(\xi, \eta)}{\mathbf{A}_\xi^R \mathbf{\Lambda}_x var(\xi)} = \frac{cov(\xi, \eta)}{var(\xi)} = \gamma. \tag{2.43}$$

When only the independent variable is latent, the regression coefficient becomes:

$$\beta = \frac{cov(F_\xi^R, \eta)}{var(F_\xi^R)} = \frac{\mathbf{A}_\xi^R \mathbf{\Lambda}_x cov(\xi, \eta)}{\mathbf{A}_\xi^R \mathbf{\Lambda}_x var(\xi)} = \frac{cov(\xi, \eta)}{var(\xi)} = \gamma. \tag{2.44}$$

When only the dependent variable is latent, the regression coefficient becomes:

$$\beta = \frac{cov(\xi, F_\eta^B)}{var(\xi)} = \frac{cov(\xi, \eta)}{var(\xi)} = \gamma. \tag{2.45}$$

2.E Regression coefficient using the bias correcting method

In this appendix, we derive the relationship between β and γ when we use the bias correcting method. To be able to perform the bias correcting method, the covariance between the true latent variable scores and the variance of the independent true latent variable scores needs to be calculated. The computation of the covariance is based on the formula given in equation (2.21):

$$cov(\xi, \eta) \quad = \quad \frac{cov(F_\xi, F_\eta)}{\mathbf{A}_\xi \mathbf{\Lambda}_x \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}}. \tag{2.46}$$

When calculating factor scores, $\mathbf{\Lambda}s$ and the $\mathbf{A}s$ matrices are readily available and the $cov(F_\xi, F_\eta)$ can be calculated. This means that it is possible to compute the true covariance. Similar calculations can be made for $cov(F_\xi, \eta)$ and $cov(\xi, F_\eta)$. The calculations for the variance of ξ is slightly more complex. First, we calculate the

variance of the factor scores:

$$\begin{aligned} var(F_\xi) &= \mathbf{A}_\xi var(\mathbf{x})\mathbf{A}_\xi^{'} \\ &= \mathbf{A}_\xi var(\mathbf{\Lambda}_x\xi + \boldsymbol{\delta})\mathbf{A}_\xi^{'} \\ &= \mathbf{A}_\xi [var(\mathbf{\Lambda}_x\xi) + var(\boldsymbol{\delta})]\mathbf{A}_\xi^{'} \\ &= \mathbf{A}_\xi [\mathbf{\Lambda}_x var(\xi)\mathbf{\Lambda}_x^{'} + \mathbf{\Theta}_{\boldsymbol{\delta}}]\mathbf{A}_\xi^{'}. \end{aligned} \tag{2.47}$$

Based on this formula, we can derive a formula for the variance of ξ:

$$\begin{aligned} \mathbf{\Lambda}_x var(\xi)\mathbf{\Lambda}_x^{'} + \mathbf{\Theta}_{\boldsymbol{\delta}} &= var(F_\xi)(\mathbf{A}_\xi\mathbf{A}_\xi^{'})^{-1} \\ \mathbf{\Lambda}_x var(\xi)\mathbf{\Lambda}_x^{'} &= var(F_\xi)(\mathbf{A}_\xi\mathbf{A}_\xi^{'})^{-1} - \mathbf{\Theta}_{\boldsymbol{\delta}} \\ \mathbf{\Lambda}_x var(\xi)\mathbf{\Lambda}_x^{'} &= (var(F_\xi) - \mathbf{A}_\xi\mathbf{\Theta}_{\boldsymbol{\delta}}\mathbf{A}_\xi^{'})(\mathbf{A}_\xi\mathbf{A}_\xi^{'})^{-1} \\ var(\xi) &= (var(F_\xi) - \mathbf{A}_\xi\mathbf{\Theta}_{\boldsymbol{\delta}}\mathbf{A}_\xi^{'})(\mathbf{A}_\xi\mathbf{\Lambda}_x\mathbf{A}_\xi^{'}\mathbf{\Lambda}_x^{'})^{-1}. \end{aligned} \tag{2.48}$$

Once the covariance and variance of the true latent variable scores are computed, the regression coefficient can be computed as $\beta = \frac{cov(\xi,\eta)}{var(\xi)}$.

3 Factor Score Path Analysis: An alternative for SEM

Abstract. Theoretical researchers consider Structural Equation Modeling (SEM) to be the preferred method to study the relationships among latent variables. However, SEM has the disadvantage of requiring a large sample size, especially if the model is complex. Furthermore, since SEM estimates all parameters simultaneously, one misspecification in the model may influence the whole model. For these reasons, researchers often use a two step Factor Score Regression (FSR) approach. In the first step, factor scores are calculated for the latent variables, which are used to perform a linear regression in the second step. However, this method results in incorrect regression coefficients. Croon (2002) developed a method that corrects for this bias. We present an extension on the method of Croon (2002), namely Factor Score Path Analysis. This method results in correct padcoefficients and has some advantages over SEM: it requires smaller sample sizes, can handle more complex models and the method is less sensitive to misspecifications, because of its stepwise nature. In conclusion, this method is a suitable alternative for SEM, when one is dealing with a complex model and small sample sizes.

This chapter has been published as Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology*, 13, 31–38. doi: 10.1027/1614-2241/a000130.

3.1 Introduction

Theoretical researchers consider Structural Equation Modeling (SEM) to be the preferred method to study the relationships among latent variables. SEM is a full information method that estimates all parameters simultaneously and results in unbiased estimates. In theory, SEM works perfectly. In practice, the method can have some drawbacks. A first issue is that SEM requires a large sample size, especially if the model is complex(Schumacker & Lomax, 1996; Valluzzi, Larson, & Miller, 2003). If the sample size is too small, the model may simply not converge and if it converges, the parameter estimations may be biased (Gagne & Hancock, 2006). A second issue originates from the simultaneous estimation of all parameters in the model. As a result, a misspecification in one part of the model may influence other parts of the model. For example, misspecifications in the structural model may bias the estimates in the measurement model.

For these reasons, limited information methods have been developed that attempt to overcome the drawbacks of SEM. For example, to overcome the misspecification issue, Bollen (1996) proposed the instrumental variables approach. This approach reduces the complexity of the model by only using one indicator per latent variable as a scaling variable. Then, the structural equations are reformulated to use this scaling variables instead of the latent variables, resulting in equations with only observed variables. These equations are finally solved using a two-stage least squares estimator. To overcome the sample size issue, applied researchers have often used the two step Factor Score Regression (FSR) approach (Lu, Kwan, Thomas, & Cedzynski, 2011). In this approach, the first step is to perform a factor analysis and to calculate factor scores for each latent variable. These factor scores are estimates for the true latent variable scores. There are several predictors that can be used to compute the factor scores, but the two most commonly used predictors in the continuous case are the regression predictor (Thomson, 1934; Thurstone, 1935) and the Bartlett predictor (Bartlett, 1937; Thomson, 1938).

In a second step, the factor scores are used in a linear regression, as if they were the true latent variable scores. By using FSR, the number of models that do not converge, is reduced. However, while this method has less problems with convergence, the use of factor scores results in biased estimates of the regression parameters, even when the sample size is large.

Several methods were developed that account for this bias. Skrondal and Laake (2001) developed a FSR method that avoids the bias by using the regression predictor for the independent latent variables and the Bartlett predictor for the dependent latent variables. However, when there are correlations between the independent variables, it turns out that this method can no longer be used. Croon (2002) developed a FSR method that corrects for the bias. This method is based on the premise that there is a difference between the variances and covariances of the factor scores and the variances and covariances of the true latent variable scores. Croon (2002) uses an estimation of the variances and covariances of the true latent variable scores, instead of the factor scores, to estimate the regression parameters. A similar approach had been discussed in Hoshino and Bentler (2013). The method of Hoshino and Bentler (2013) only uses the Bartlett predictor and relies on weighted least squares (WLS) estimation, while the method of Croon (2002) can be used with any predictor and any estimator. In his paper, Croon (2002) only discussed the method from a population point of view and did not study how the method performs in finite samples. This was done by Lu et al. (2011) and by Devlieger, Mayer, and Rosseel (2016). Both concluded that the method of Croon (2002) results in unbiased parameter estimates when finite samples are used. Devlieger et al. (2016) also showed that the method has a comparable efficiency, mean square error, power and type I-error rate as SEM, when the sample size is large. However, despite these encouraging results, many questions about this method still remain. For example, how does the method compare to SEM with regard to it's main drawbacks, namely settings with small sample sizes and misspeci-

fications? And can the method be extended to be used with path analysis? The goal of this paper is to answer these questions.
The rest of the paper is organized as follows. First, we outline the method of Croon, including a step by step description of how to perform the method using path analysis. Next, two simulation studies will be presented. In these studies, the performance of the method of Croon will be compared to the performance of SEM. The first study uses a correctly specified and two incorrectly specified models. The goal of this study is to evaluate the finite sample performance of the method of Croon using path analysis and to compare the robustness of SEM and the method of Croon to misspecifications in the model. The second study uses a more complex, but correctly specified model, and smaller sample sizes. The aim of this study is to compare the performance of both methods when the sample size is small.

3.2 The method of Croon

3.2.1 The method of Croon using regression analysis

Croon (2002) developed the method to be used in general latent variable models, meaning all models that include latent variables. He uses univariate and multivariate regression settings with several latent variables to explain the method. The simple regression model in Figure 3.1 is an example of this setting.
The first step of the method is to use the measurement models to perform a factor analysis for each latent variable separately and to calculate their respective factor scores, F_ξ and F_η. As we mentioned before, there are several predictors that can be used to compute these factor scores, such as the regression predictor and the Bartlett predictor. In this paper, the regression predictor will be used.
The second step of FSR methods is to perform a linear regression between the factor scores, resulting in a regression coefficient. In a simple linear regression, the true regression coefficient is defined as the true covariance between the dependent and the independent

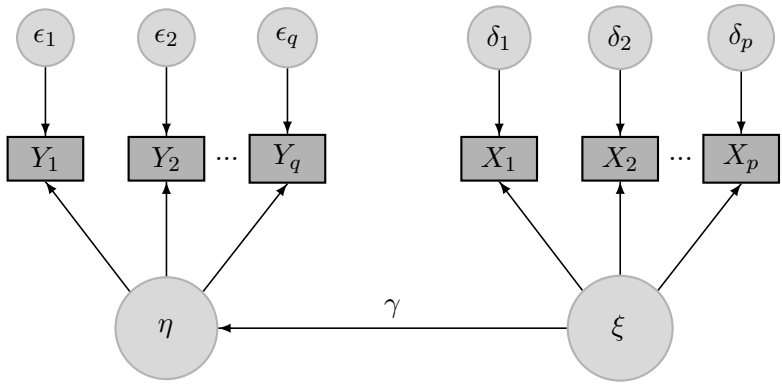


Figure 3.1 The simple regression model

variable, multiplied by the inverse of the true variance of the independent variable:

$$\gamma = cov(\xi, \eta) var(\xi)^{-1}. \tag{3.1}$$

When performing the linear regression with factor scores, the regression coefficient is defined as the covariance between the factor scores of the dependent and the independent variable, multiplied by the inverse of the variance of the factor scores of independent variable:

$$\beta = cov(F_\xi, F_\eta) var(F_\xi)^{-1}. \tag{3.2}$$

Croon (2002) has shown that γ and β are not equal in all conditions, since there is a difference between the variances and covariances of the factor scores ($\mathbf{\Sigma}_{FS}$) and the variances and covariances of the true latent variable scores ($\hat{\mathbf{\Sigma}}_\eta$). For this reason, Croon (2002) uses estimates of the variances and covariances of the true latent variable scores ($\hat{\mathbf{\Sigma}}_\eta$) instead of these of the factor scores in his method. The variances and covariances of the true latent variable scores can be estimated as follows:

$$\widehat{cov(\xi, \eta)} = cov(F_\xi, F_\eta)(\mathbf{A}_\xi \mathbf{\Lambda}_x \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'})^{-1}, \tag{3.3}$$

$$\widehat{var(\xi)} = (var(F_\xi) - \mathbf{A}_\xi \mathbf{\Theta}_{\boldsymbol{\delta}} \mathbf{A}_\xi^{'})(\mathbf{A}_\xi \mathbf{\Lambda}_x \mathbf{A}_\xi^{'} \mathbf{\Lambda}_x^{'})^{-1}, \tag{3.4}$$

with, $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ the factor loadings, $\mathbf{A}_\xi$ and $\mathbf{A}_\eta$ the factor scores matrices and $\mathbf{\Theta}_{\boldsymbol{\delta}}$ the covariance matrix of $\boldsymbol{\delta}$, with $\boldsymbol{\delta}$ the vector of measurement errors associated with the indicators $X_1, X_2, ..., X_p$ of ξ. Once the covariance and variance of the true latent variable scores are computed, the regression coefficient can be computed as

$$\hat{\beta} = \widehat{cov(\xi, \eta)}\widehat{var(\xi)}^{-1}.$$

For more details on how these formulas were derived, the interested reader is referred to (Croon, 2002) and (Devlieger et al., 2016).

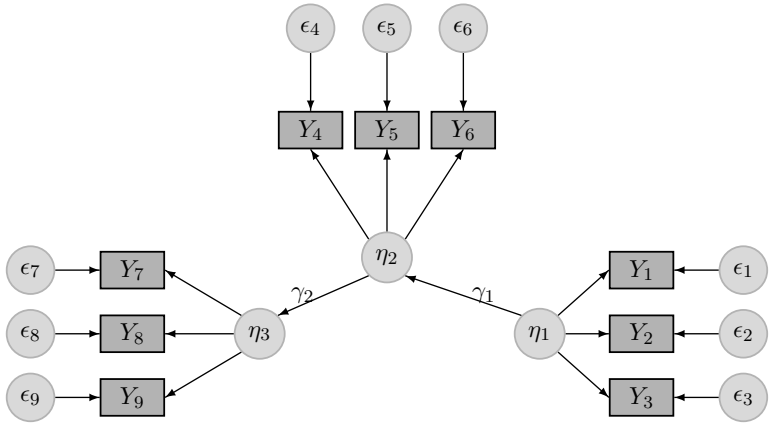
3.2.2 The method of Croon using path analysis

When the model includes a mediational relationship, it is not possible to perform one single linear regression. For non-recursive models such as the full mediation model in Figure 3.2a or the partial mediation model in Figure 3.2b, it is possible to perform a series of linear regression analyses, one per endogeneous variable. The full mediation model could also be analysed using the method of Skrondal and Laake (2001), while the method would fail for the partial mediation model, because of the correlation between η_1 and η_2. However, this multiple regression strategy can only be used for recursive path models. For nonrecursive path models (having reciprocal effects, loops, or bow-pattern disturbance correlations), a simultaneous estimation method like maximum likelihood (ML) is required (Kline, 2015). Therefore, we apply the principle of the method of Croon to path analysis. When using path analysis, the method can be summarized as follows:

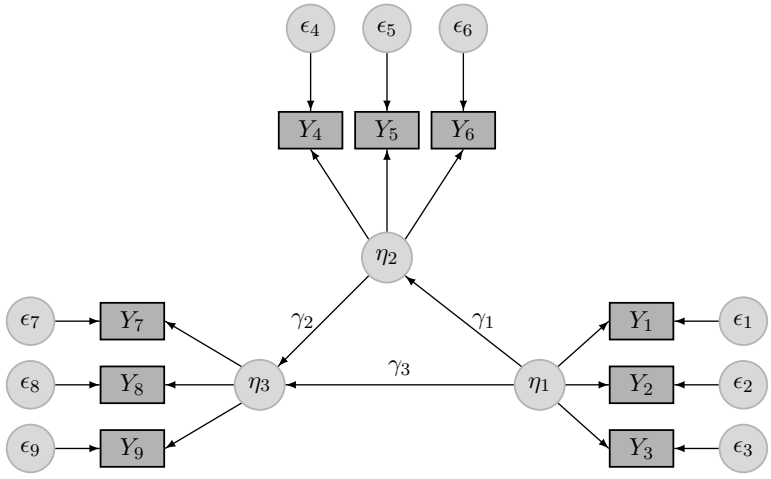
1. Perform factor analysis for all latent variables separately and calculate their respective factor scores.

2. Calculate the variance-covariance matrix of the factor scores ($\mathbf{\Sigma}_{FS}$).

3. Estimate the true variances and covariances for all elements in this variance-covariance matrix ($\hat{\mathbf{\Sigma}}_\eta$).



(a) A full mediation model



(b) A partial mediation model

Figure 3.2 Different kinds of mediation models

4. Perform a path analysis, using the estimated variances and covariances $\hat{\boldsymbol{\Sigma}}_\eta$ as the input covariance matrix for the model.

Combining the method of Croon and path analysis means nonrecursive models can be analysed using factor scores, without bias. An example, using a bow pattern model, is given in Figure 3.3.

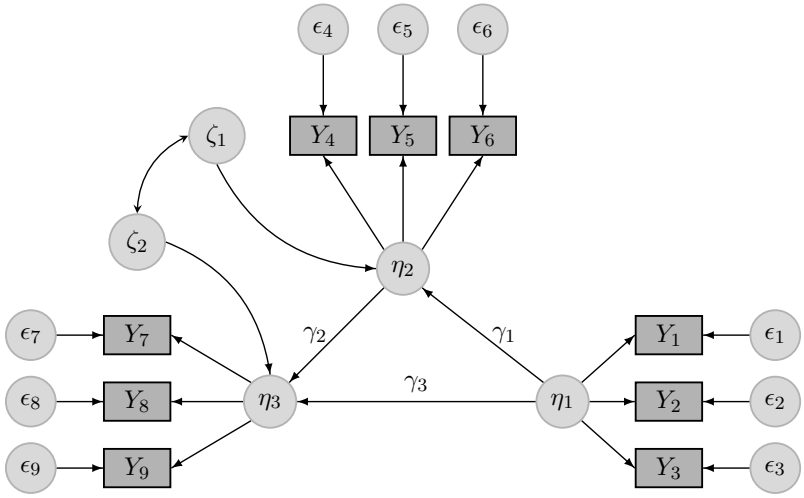
3.3 Simulation studies

3.3.1 Study 1: Path analysis and misspecifications

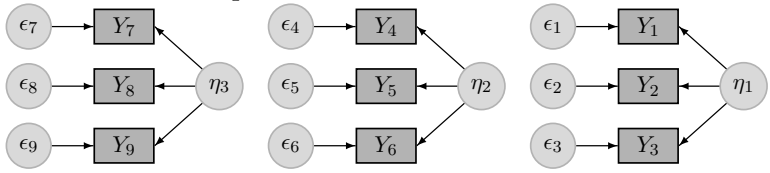
Data simulation

The data of the first study was simulated using the ground truth model depicted in Figure 3.4. The data was simulated in R (R Development Core Team, 2016). The true latent variable scores were generated first. The variances of the exogeneous variables ξ_1, ξ_2 and ξ_3 were set at 100, while the residual variances of the endogeneous variables η_1 and η_2 were both set at 400. The true latent scores of ξ_1, ξ_2 and ξ_3 were generated first, followed by the regression residuals ζ_1 and ζ_2, all by drawing from a univariate normal distribution. Finally, using the structural equations $\eta_1 = \gamma_1\xi_1 + \gamma_2\xi_2 + \zeta_1$ and $\eta_2 = \gamma_3\xi_3 + \gamma_4\xi_1 + \gamma_5\eta_1 + \zeta_2$, the true latent scores on η_1 and η_2 were generated. γ_1, γ_2, γ_3 and γ_4 were set at 1.5, while γ_5 was set at a value of 0.51.

Then, data for each observed item response x_{lm} and y_{lm} was generated, with the 'l' index referring to the latent variables and the 'm' index to the items. The measurement models of the latent variables were $y_{lm} = \lambda_{y_{lm}}\eta_l + \epsilon_{lm}$ and $x_{lm} = \lambda_{x_{lm}}\xi_l + \delta_{lm}$. All factor loadings were set at 1. The residual variances were set at $\Theta_{\epsilon_{lm}} = \frac{var(\eta_l)(1-CD_{y_l})}{CD_{y_l}}$ and $\Theta_{\delta_{lm}} = \frac{var(\xi_l)(1-CD_{x_l})}{CD_{x_l}}$, with CD_{y_l} and CD_{x_l} the coefficients of determination for the measurement models, respectively. All CD_{x_l} and CD_{y_l} are equal and will thus be referred to as CD. The coefficients of determination CD were varied between 0.3, 0.6, 0.7 and 0.9. The sample size was set at 500, 1000 or 2000. Together, this created 12 experimental conditions.



1. Perform factor analysis for all latent variables separately and calculate their respective factor scores.



2. Calculate the variance-covariance matrix of the factor scores ($\mathbf{\Sigma}_{FS}$).

3. Estimate the true variances and covariances for all elements in this variance-covariance matrix ($\hat{\mathbf{\Sigma}}_{\eta}$).

4. Perform a path analysis, using the estimated variances and covariances $\hat{\mathbf{\Sigma}}_{\eta}$ as the input covariance matrix for the model.

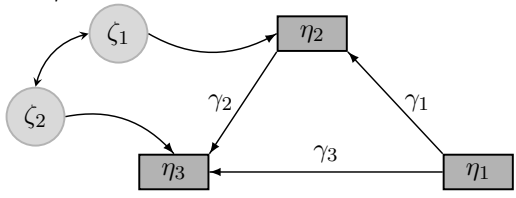


Figure 3.3 The method of Croon using path analysis, for a bow pattern model

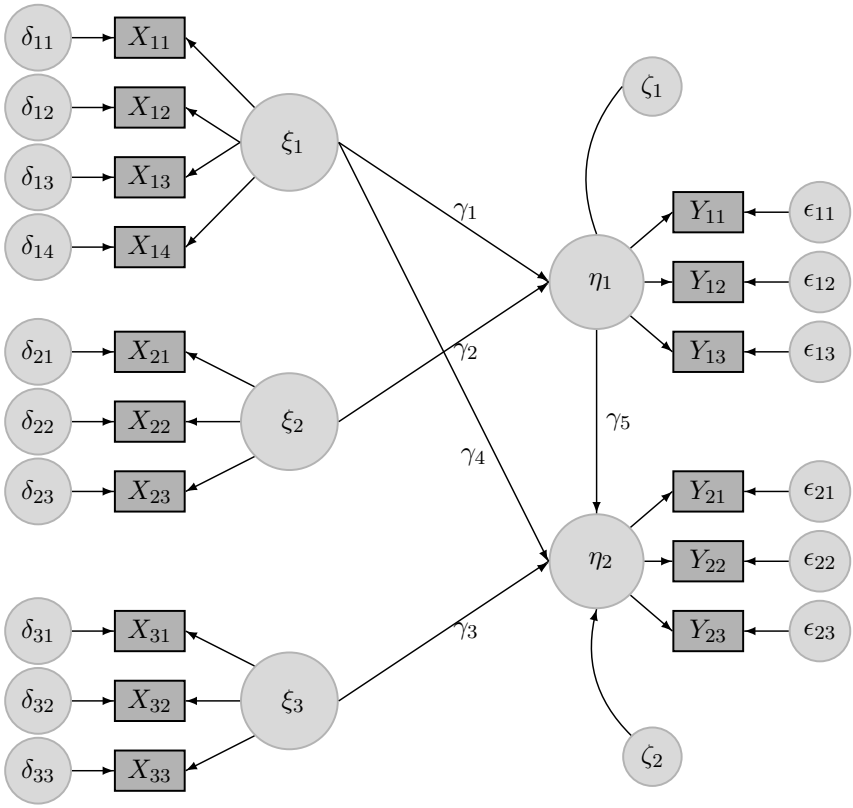


Figure 3.4 The ground truth of Study 1

Analyses

For each of the 12 conditions, 1000 datasets were generated and then analysed with both the method of Croon and SEM, using a "maximum likelihood" estimator. For the SEM approach, lavaan (Rosseel, 2012) was used. For the Croon method, lavaan (Rosseel, 2012) was used to calculate the factor scores and our own written routines were used to compute the Croon corrections ans the resulting regression coefficients of the structural part of the model. For both methods, three different models were fitted to the data. The first model was correctly specified, creating a condition where SEM works optimally (Model 1). The two other models were misspecified models, one misspecification in the structural model and one misspecification in the measurement model. For the measurement misspecification (Model 2), the item X_4 was set to measure ξ_2 instead of ξ_1. For the structural misspecification (Model 3), the regression parameter γ_4 was fixed to 0. This means there is no longer a direct effect of ξ_1 on η_2, but there is still a mediated effect through η_1. For each model, the five regression coefficients γ_i were obtained for both methods. Based on the 1000 replications, two performance criteria were computed, namely the convergence rate and the bias of each regression coefficient.

Results

Convergence rate

The convergence rate for all three models is depicted in Figure 3.5. When the model is correctly specified or when there is a structural misspecification, both the method of Croon and SEM had a convergence rate of 100 % in every condition. When there is a measurement misspecification, the convergence rate drops for both methods, but more severely for SEM. The convergence rate ranges from 0.701 to 0.911 for the method of Croon, while it ranges from 0.386 to 0.827 for SEM.

In conclusion, the method of Croon does indeed converge more often than SEM, when there are measurement misspecifications in the

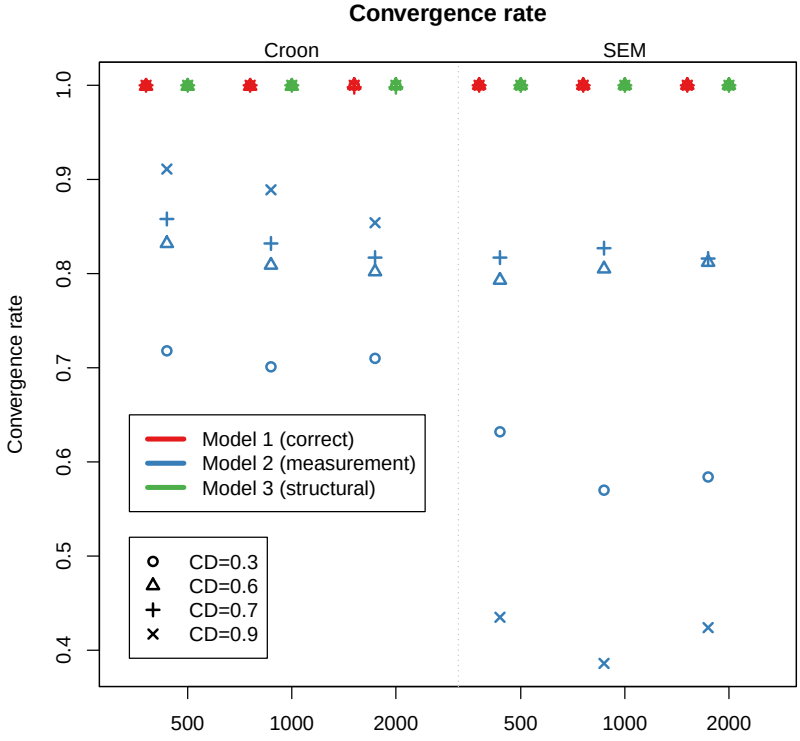


Figure 3.5 The convergence rate for SEM and the Croon method for the four models

model. It is important to know if the method not only converges, but also results in reliable parameter estimates. For this reason, we also study the bias.

Bias When the model is correctly specified, both methods show almost no bias (Figure 3.6). However, there is a difference between the two methods. SEM tends to slightly overestimate, while the method of Croon tends to underestimate the regression coefficient. The small bias that can be found, disappears with growing sample size and coefficient of determination.

When looking at Figure 3.7, it is clear that there is a huge bias in

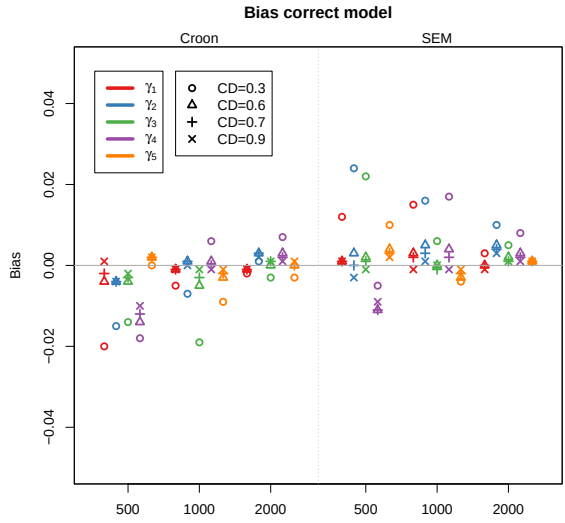


Figure 3.6 The bias for SEM and the method of Croon for the correctly specified Model 1. Note the scale of the y-axis.

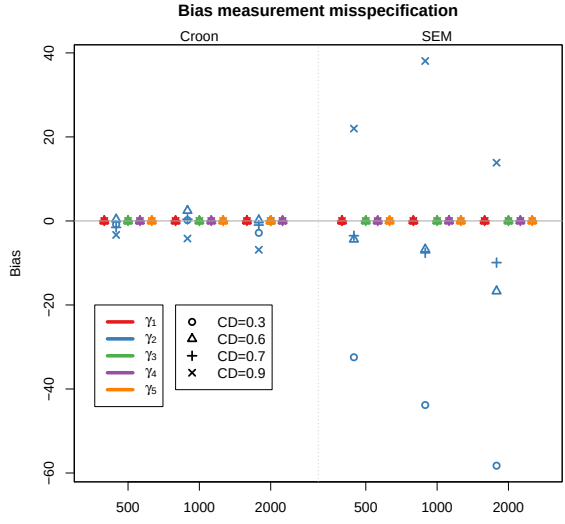


Figure 3.7 The bias for SEM and the method of Croon for Model 2, which has a misspecification in the measurement model

regression parameter γ_2 when the measurement model is misspecified. This is true for both methods, but especially for SEM. Only the parameter estimate that is directly related to the misspecified latent variable ξ_2 is affected. This is the latent variable that was measured by an item that does not really measure this construct. The latent variable ξ_1, that was measured by one item less than in the ground truth, is unaffected.
For the structural misspecification, the regression parameter that was set to 0, γ_4, was excluded from the graphs (Figure 3.8). The Croon method only shows bias in parameter γ_5, which is to be expected given the ground truth model in Figure 3.4. For SEM, all regression parameters are biased.
It can be concluded that SEM is less robust against misspecifications. SEM converges considerably less than the Croon method if there is a measurement misspecification. When the model does converge, there is also more bias in the estimates of the regression parameters, both for structural and measurement mispecifications.

3.3.2 Study 2: Small sample size

Data simulation

For this study, a more complex model was used, with three exogeneous and three endogeneous variables. The model is depicted in Figure 3.9. The data was generated in the same manner as in Study 1. The variances of the exogeneous variables ξ_1, ξ_2 and ξ_3 were set at 100, while the residual variances of the endogeneous variables η_1, η_2 and η_3 were set at 400. Different sample sizes, namely 50,100 200 and 300 were used. The coefficient of determination was also varied, namely 0.3, 0.6, 0.7 and 0.9. Combined, this gave 16 conditions.

Analyses

For each condition, 1000 datasets were generated and then fitted with a correctly specified model, using both SEM and the method of

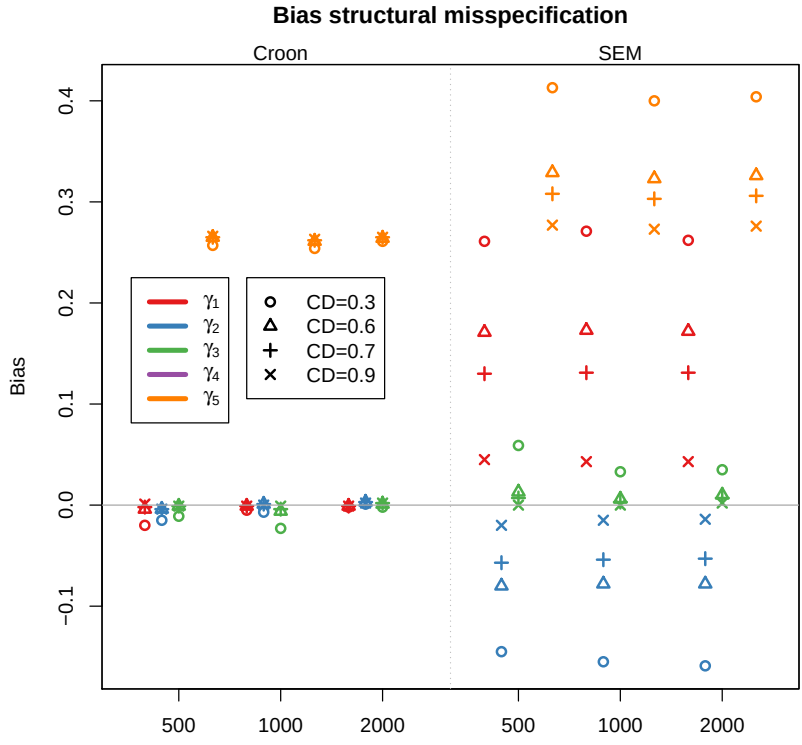


Figure 3.8 The bias for SEM and the method of Croon for Model 3, which has a misspecification in the structural model.

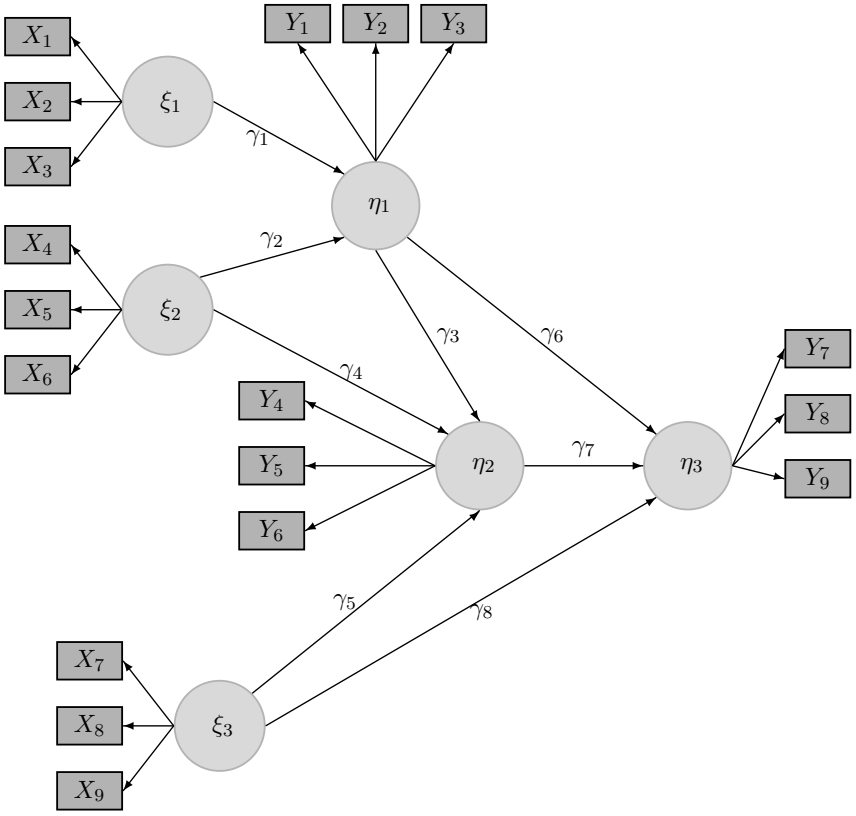


Figure 3.9 The ground truth of Study 2

Croon. Then, two criteria were computed, namely the convergence rate and the bias in the regression parameters.

Results

Convergence rate The first criterium is the convergence rate. As can be seen in Figure 3.10, the proportion is higher for the Croon method than for SEM. The lowest proportion for the method of Croon is 0.92 and there are only two conditions in which the method does not converge every time. These are the conditions with very weak factor loadings (CD=0.3) and a very small sample size (50 or 100). The lowest proportion for SEM is 0.752. The proportion

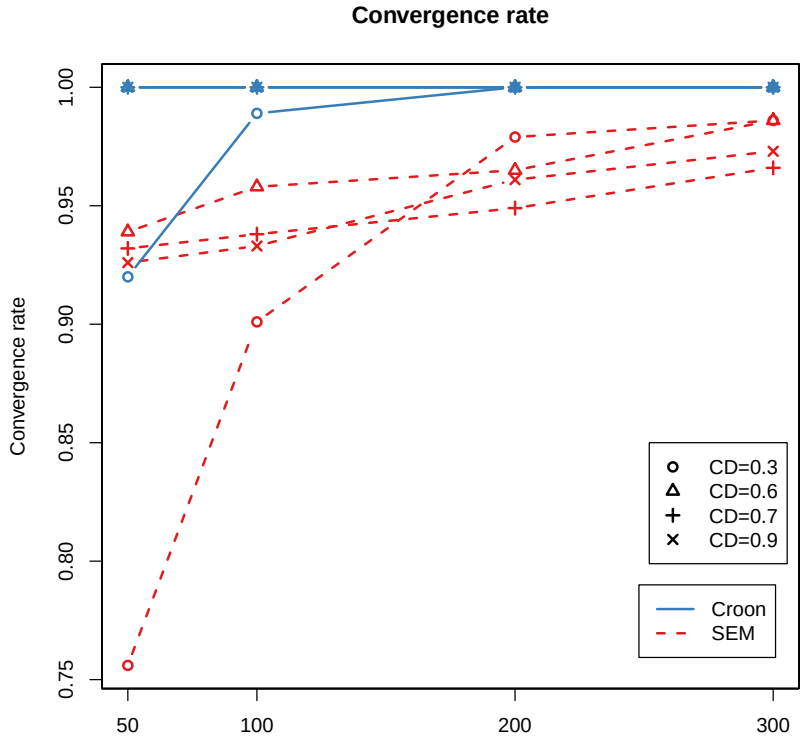


Figure 3.10 The convergence rate for SEM and the method of Croon, using the model as depicted in Figure 3.9.

does increase as the sample size and coefficient of determination increases, but the proportion does not reach 1, even when the sample size is 300. To summarize, the method of Croon converges more often than SEM when the model is complex and the sample size is small.

Bias

The second criterium is the bias in the regression parameters. Since there are three endogeneous variables, the regression parameters were divided into three groups, namely γ_{1-2}, γ_{3-5} and γ_{6-8}. The results are different for the three groups (see Figure 3.11).

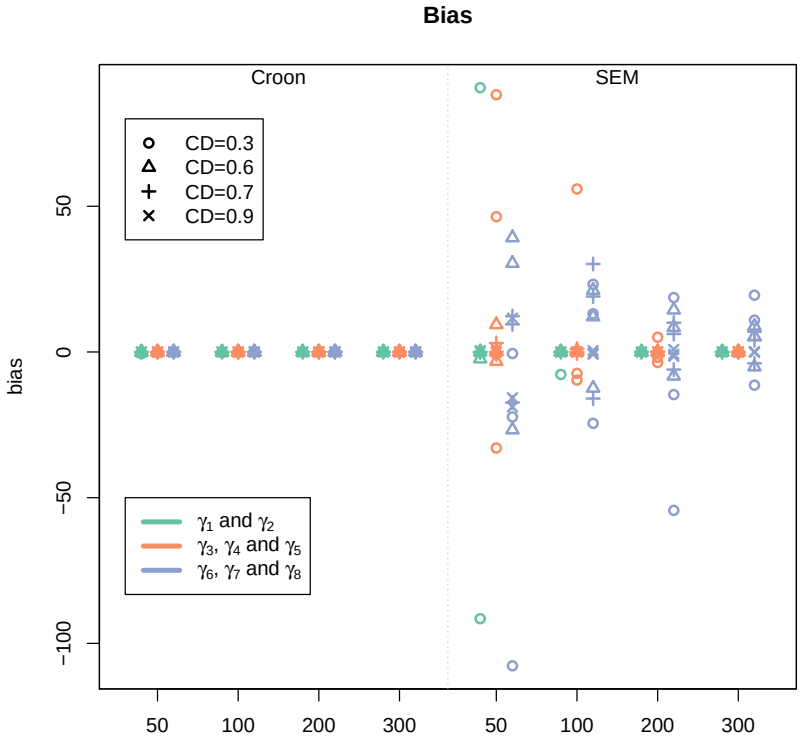


Figure 3.11 The bias of both methods per sample size and coefficient of determination

For the Croon method, there is almost no bias in all conditions for all three groups. For the SEM method, different patterns can be detected. The first group of parameter estimates, γ_{1-2}, only shows bias when the factor loadings are weak (CD=0.3) and the sample size is lower than 200. The second group of parameter estimates, γ_{3-5}, shows bias when the factor loadings are weak (CD=0.3) and the sample size is lower than 300. The third group of parameter estimates, γ_{6-8}, shows bias in almost all conditions. There is only no bias when the factor loadings are strong (CD=0.9) and the sample size is higher than 50. The difference between the three groups can be explained by looking at the model in Figure 3.9. The first endogeneous variable η_1 is only directly influenced by exogeneous variables, while the second endogeneous variable η_2 is also influenced by another endogeneous variable (η_1) and the third endogeneous variable η_3 is influenced by both η_1 and η_2. When combining this information with the results regarding the bias, one could assume that the more complex the model gets, the larger the sample size needs to be to be able to get unbiased parameter estimates.

In conclusion, the method of Croon is indeed a better alternative than SEM when the sample size is rather small or the model rather complex. The method of Croon has less problems to converge correctly when the sample size is small and gives less biased estimates of the regression parameter when the model is complex.

3.4 Discussion

We compared the method of Croon, a factor score regression method, to SEM using path analysis. The two studies gave us an overall comparison between the performance of SEM and the method of Croon, with regard to the bias and convergence rate. We showed that the method of Croon performs just as well as SEM with regard to bias and convergence rate when path analysis is used. It also handles misspecifications better than SEM and requires a smaller sample size. It can be concluded that the method

of Croon is a suitable alternative for SEM when the researcher prefers factor score regression over SEM or when the model does not converge using SEM.

However, there are some settings in which SEM is still the best alternative. Unless additional restrictions are implemented, factor score regression methods only work when there are at least three items per latent variable. For the moment, the method of Croon also doesn't work for connected measurement models, such as models with cross-loadings or correlated residual errors. However, we are currently extending the method to be able to handle these kinds of models. We also want to make some extensions to the inference of the model. In future research, we want to use standard errors that are suitable for two-step estimation methods, and we want to develop fit indices, so that the fit of the model can be evaluated, just as in SEM. Finally, we also plan to implement the method in lavaan (Rosseel, 2012).

References

Bartlett, M. (1937, jul). The statistical conception of mental factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Bollen, K. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109–121. doi: 10.1007/BF02296961

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi: 10.1177/0013164415607618

Gagne, P., & Hancock, G. R. (2006). Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. *Multivariate Behavioral Research*, *41*(1), 55–64. doi: 10.1207/s15327906mbr4101

Hoshino, T., & Bentler, P. M. (2013). Bias in Factor Score Regression and a Simple Solution. In A. R. de Leon & K. C. Chough (Eds.), *Analysis of mixed data- methods & applications* (pp. 43–61). Chapman and Hall. doi: 10.1201/b14571-5

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (Fourth ed.). Guilford publications.

Lu, I. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011, sep). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, *28*(3), 258–268. doi: 10.1016/j.ijresmar.2011.03.006

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical computing.

Rosseel, Y. (2012). lavaan : An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Schumacker, R., & Lomax, R. (1996). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. doi: 10.1007/BF02296196

Thomson, G. (1934). The meaning of i in the estimate of g. *British Journal of Psychology*, *25*, 92–99. doi: 10.1111/j.2044-8295.1934.tb00728.x

Thomson, G. (1938, feb). Methods of Estimating Mental Factors. *Nature*, *141*(3562), 246–246. doi: 10.1038/141246a0

Thurstone, L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press. doi: 10.1037/h0075959

Valluzzi, J. L., Larson, S. L., & Miller, G. E. (2003). Indications

and Limitations of Structural Equation Modeling in Complex Surveys: Implications for an Application in the Medical Expenditure Panel Survey (MEPS). In *Joint statistical meetings - section on survey research methods indications* (pp. 4345–4352). doi: 10.1.1.493.467

4 New developments in FSR: fit indices and a model comparison test

Abstract. Factor score regression (FSR) is a popular alternative for structural equation modeling (SEM). Naively applying FSR induces bias for the estimators of the regression coefficients. Croon (2002) proposed a method to correct for this bias. Next to estimating effects without bias, interest often lies in inference of regression coefficients or in the fit of the model. In this paper we propose fit indices for FSR that can be used to inspect the model fit. We also introduce a model comparison test based on one of these newly proposed fit indices that can be used for inference of the estimators on the regression coefficients. In a simulation study we compare FSR with Croon's corrections and SEM in terms of bias of the regression coefficients, type I error rate and power.

This chapter has been published as Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New Developments in Factor Score Regression: Fit Indices and a Model Comparison Test *Educational and Psychological Measurement*, doi: 10.1177/0013164419844552.

4.1 Introduction

In behavioral and social sciences, interest often lies in relationships between latent variables. Typical examples of such variables are motivation, self-confidence or positive mood. When the data are continuous, structural equation modeling (SEM), using maximum likelihood estimation (MLE), is often considered to be the best method to analyze the relationships between these latent variables (Bentler & Chou, 1987; Jöreskog, 1973). However, because MLE estimates all parameters simultaneously, it is sensitive to misspecifications in the model. Misspecification in one part of the model (for instance the structural model) may affect parameter estimates in other parts of the model (for instance the measurement model) (Bollen, 1996). A second limitation of MLE are the sample size requirements (McNeish, 2016; Nevitt & Hancock, 2004; Schumacker & Lomax, 1996). For example, when the sample size is smaller than the number of observed variables, MLE will not be able to estimate any parameters, unless we use a casewise likelihood approach. When the sample size is larger than the number of observations, small sample sizes can lead to non-convergence, bias in the parameters and a lack of power (Gagne & Hancock, 2006). Sample size can be a serious issue for applied researchers, since a large sample size is often impossible due to financial, practical or time restrictions.

These two shortcomings are well known and different solutions have been proposed, for example two-step SEM (Anderson & Gerbing, 1988; Bakk & Kuha, 2017), penalized likelihood estimation (Huang, Chen, & Weng, 2017; Jacobucci, Grimm, Brandmaier, Serang, & Kievit, 2018), the instrumental variable approach (Bollen, 1996) or using a Bayesian estimation method (Scheines, Hoijtink, & Boomsma, 1999). In this paper we will focus on a stepwise alternative called factor score regression (FSR). FSR proceeds as follows: First, a factor analysis is conducted to predict scores on latent variables, hereafter referred to as factor scores. For measurement models with simple structure, a factor model is fitted for each

latent variable separately. This prevents possible misspecifications to spread to other parts of the model. In addition, because only a small subset of the parameters is involved for each latent variable, estimation is much more stable. Measurement models are only jointly estimated when they are linked (by correlated measurement errors, crossloadings or equality constraints), or when data is incomplete, and more variables are needed to assure the MAR assumption is valid. In the second step, linear regression or path analysis is used to estimate the regression coefficients between factor scores and possibly observed variables (Lu, Thomas, & Zumbo, 2005). It has been shown that using factor scores in this second step induces bias in the estimators of the regression coefficients (e.g. Bollen, 1989; Lastovicka & Thamodaran, 1991; Lewis, 2005; Shevlin, Miles, & Bunting, 1997). Several methods have been developed to correct for this bias. Croon (2002) corrects for the bias by adapting the variance-covariance matrix of the factor scores, while Skrondal and Laake (2001) avoids the bias by using two different methods to predict the factor scores. As Devlieger and Rosseel (2017) point out, the method of Skrondal and Laake does not allow exogenous variables to be correlated. Therefore we will focus on the bias correcting method of Croon.

Devlieger, Mayer, and Rosseel (2016) and Lu, Kwan, Thomas, and Cedzynski (2011) showed that FSR with Croon's corrections and MLE perform equally well in terms of bias if the sample size is large. When the sample size is small, FSR with Croon's corrections even outperforms MLE, both in terms of bias and convergence rate (Devlieger & Rosseel, 2017; Lu et al., 2011). In addition, the Croon method is more robust against misspecifications both in the structural part and in the measurement part of the model (Devlieger & Rosseel, 2017). Recently, it has been shown that the Croon method also outperforms MLE in the multilevel setting, in particular when the number of clusters at the between level is rather small (Devlieger & Rosseel, 2019).

Despite these positive results, the method still has its shortcomings.

In the literature, focus mainly lies on estimating the regression parameters between variables without bias. How to draw inference, perform model comparisons and determine fit indices are yet to be further explored. There are some ways to draw inference, such as the two-step standard error developed by Rosseel (2019). However, this focuses on single parameter tests. How to draw inference for multiple parameters, such as in an omnibus test for categorical covariates or a model comparison test, has not been studied yet. When using MLE, a model comparison test is typically based on a 'Likelihood Ratio Test (LRT)', which in the SEM context results in the χ^2-difference test. This χ^2-fit index, or other fit indices, are currently unavailable for the factor score regression method.

In this paper, we propose a newly developed fit index using FSR, χ^2_a, which is an approximation for χ^2. Based on this χ^2_a, several other approximated fit indices can be calculated. This means that the global fit of the model can now also be evaluated using FSR. The newly proposed χ^2_a can also be used to conduct a model comparison test when the restricted model is nested in the full model.

This paper is organized as follows. We start by introducing structural equation models, how these can be estimated using MLE and how the global fit can be evaluated. Next, we show how a structural equation model can be estimated by the method of Croon, and propose the following extensions. First, we introduce fit indices for FSR to inspect the model fit. Secondly, we use one of these fit indices to conduct a model comparison test, both for continuous and categorical predictors. To evaluate these newly developed extensions, we conduct a simulation study where we compare the performance of SEM and FSR in terms of bias, type I error rate, power and fit indices. Finally, we illustrate our findings, using a real-world dataset.

4.2 Structural equation modeling

4.2.1 The model

A structural equation model often contains a structural and a measurement part. The structural part models the relationship between the latent variables, while the measurement part indicates how the latent variables are measured. Ignoring the mean structure, a structural equation model can be written as:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta} \tag{4.1}$$

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{4.2}$$

with $\mathbf{B}$ a matrix of regression coefficients, $\boldsymbol{\zeta}$ a vector of the residual error terms, $\boldsymbol{\Lambda}$ the matrix of factor loadings of the observed variables on the latent variables, $\mathbf{y}$ a vector of indicators measuring the latent variables $\boldsymbol{\eta}$, and $\boldsymbol{\epsilon}$ a vector of measurement error variables. If we assume $\text{cov}(\boldsymbol{\eta}, \boldsymbol{\zeta}) = \mathbf{0}$ and if we write $\text{Var}(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$, it follows that

$$\boldsymbol{\Sigma}_{\eta} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})^{-1'}, \tag{4.3}$$

with $\mathbf{I}$ the identity matrix. If we further assume $\text{cov}(\boldsymbol{\eta}, \boldsymbol{\epsilon}) = \mathbf{0}$ and write $\text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Theta}$, it follows that

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\eta}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \tag{4.4}$$

4.2.2 Estimation using MLE

Different methods can be used to estimate the parameters of a structural equation model, such as Maximum Likelihood Estimation (MLE), Weighted Least Squares (WLS) and Generalized Least Squares (GLS). When $\mathbf{y}$ is continuous, MLE is the most commonly used method in SEM (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Therefore, we will focus on MLE in this paper. Using MLE, all parameters are estimated simultaneously by minimizing the discrepancy function

$$\mathrm{F}_{ML} = tr(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}) + log|\hat{\mathbf{\Sigma}}| - log|\mathbf{S}| - k, \tag{4.5}$$

with $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of free parameters, k the number of observed variables and $\mathbf{S}$ the observed variance-covariance matrix. Note that to be able to minimize this function, $\mathbf{S}$ has to be positive-definite, implying that the sample size cannot be smaller than the number of observed variables.

4.2.3 Fit indices

To examine the fit between the model and the data, a wide range of fit indices is used in SEM (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). Most fit indices are a function of the estimated model implied variance-covariance matrix $\hat{\mathbf{\Sigma}}$ and the observed variance-covariance matrix $\mathbf{S}$ of the observed variables (Schermelleh-Engel et al., 2003). $\mathbf{S}$ is computed based on the data, while $\hat{\mathbf{\Sigma}}$ is estimated based on the parameter estimates of the model (Kaplan, 1955), using formula (4.4).

We will discuss four of the most commonly used indices, namely the χ^2-test statistic, the root mean square error of approximation (RMSEA), the standardised root mean squared residual (SRMR) and the Comparative Fit Index (CFI). The significance test based on the χ^2-test statistic yields a decision about the fit of the model and is well known. However, the χ^2-value is highly dependent on the complexity of the model and on the sample size. Schermelleh-Engel et al. (2003) states that for these reasons, one should not place too much importance on the significance of this test. The other fit indices are all rather descriptive. For each of the descriptive fit indices, different cut-off criteria can be found in the literature. Note that these cut-off criteria are just rules of thumb and that the different fit indices can result in contradicting conclusions concerning the model fit. Therefore, it is important to use several fit indices simultaneously. Recently, significance tests for RMSEA and SRMR were developed (Maydeu-Olivares, 2017).

The χ^2-test statistic is calculated by multiplying the minimum

fit function F_{ML} with the sample size n:

$$\chi^2 = n\mathrm{F}_{ML}, \tag{4.6}$$

Under the null hypothesis that the observed variance-covariance matrix **S** is equal to the model-implied variance-covariance matrix $\hat{\mathbf{\Sigma}}$, χ^2 follows a χ^2-distribution with df = s-t degrees of freedom. s is the number of non-redundant elements in the vector of sample statistics (here **S**) and t the number of parameters to be estimated. If the p-value corresponding with the χ^2-value is larger than the significance level, the null hypothesis is accepted and the model is considered to fit the data well.

RMSEA is based on χ^2 and quantifies the approximation error due to model misspecification , taking the sample size and degrees of freedom into account:

$$RMSEA = \sqrt{max(\frac{\chi^2 - \mathrm{df}}{\mathrm{df}(n-1)}, 0)} \tag{4.7}$$

A value of 0.06 or below is often considered to indicate a good fit between the model and the data (Hu & Bentler, 1999).

SRMR is not based on χ^2, but instead uses the residuals, or the differences between the elements of $\hat{\mathbf{\Sigma}}$ and **S**, scaled by the standard deviations of the observed variables (= square root of the diagonal elements of **S**).

$$SRMR = \sqrt{\frac{\sum_{i=1}\sum_{j=1}[(s_{ij} - \hat{\sigma}_{ij})/(s_i s_j)]^2}{k(k+1)/2}}, \tag{4.8}$$

with s_{ij} an element of **S**, $\hat{\sigma}_{ij}$ an element of $\hat{\mathbf{\Sigma}}$, $s_i = \sqrt{s_{ii}}$, $s_j = \sqrt{s_{jj}}$ and k the number of observed variables. A value of 0.08 or below is often considered to indicate a good fit between the model and the data (Hu & Bentler, 1999).

CFI is a model comparison between the model of interest and a baseline model. Often, the independence model is used as the baseline model (Schermelleh-Engel et al., 2003). The independence model assumes that all variables are uncorrelated. In this model

only the variances of the observed variables have to be estimated. In MLE, the implied variance-covariance matrix of the baseline model $\hat{\mathbf{\Sigma}}_0$ is thus a diagonal matrix with the diagonal elements equal to the variances of the observed variables[1]. Based on $\hat{\mathbf{\Sigma}}_0$, the χ^2-test statistic can also be calculated for the baseline model:

$$\begin{aligned} \mathrm{F}_{ML_0} &= tr(\mathbf{S}\hat{\mathbf{\Sigma}}_0^{-1}) + log|\hat{\mathbf{\Sigma}}_0| - log|\mathbf{S}| - k \qquad (4.9) \\ \chi_0^2 &= n\mathrm{F}_{ML_0} \qquad (4.10) \end{aligned}$$

χ_0^2 follows a χ^2-distribution with $\mathrm{df}_0 = \frac{k(k-1)}{2}$ degrees of freedom.

$$CFI = 1 - \frac{max[(\chi^2 - \mathrm{df}), 0]}{max[(\chi_0^2 - \mathrm{df}_0), 0]} \qquad (4.11)$$

A value of 0.95 or larger is often considered to indicate a good fit between the model and the data (Hu & Bentler, 1999).

4.3 Factor score regression with Croon's corrections

4.3.1 Estimation

Instead of estimating all parameters simultaneously, applied researchers often use factor score regression, which estimates the structural and measurement parts of the model in two separate steps. The measurement part is estimated first by performing a factor analysis for each latent variable. Based on these factor analyses, factor scores are calculated. The structural part is then estimated next by using these factor scores in a linear regression or a path analysis.

Using the factor scores naively causes bias, because the variance-covariance matrix of the factor scores is not the same as the

[1]Note that for other estimators (GLS, WLS), the estimated values of the "variances" of the observed variables in this independence model do **not** necessarily correspond to the observed diagonal elements of **S**.

variance-covariance matrix of the true latent variable scores. Croon (2002) developed formulas that can correct the variance-covariance matrix. For more details on these correction formulas, the interested reader is referred to Croon (2002) and Devlieger et al. (2016). To conduct a factor score regression with Croon's corrections, we propose the procedure as presented by Devlieger and Rosseel (2017).

1. Conduct a separate factor analysis for each latent variable and predict factor scores.

2. Calculate the variance-covariance matrix of the predicted factor scores ($\mathbf{S}_F$).

3. Correct the variance-covariance matrix $\mathbf{S}_F$, using the formulas of Croon (Croon, 2002) to obtain an estimator for the true variance-covariance matrix ($\hat{\mathbf{\Sigma}}_\eta$).

4. Perform a path analysis or a linear regression with predicted factor scores using the estimator for the true variance-covariance matrix $\hat{\mathbf{\Sigma}}_\eta$.

4.3.2 Fit indices using the method of Croon

Global fit indices

We have already discussed four fit indices that are commonly used in SEM. All four indices are based on $\hat{\mathbf{\Sigma}}$ and $\mathbf{S}$. To be able to determine fit indices for the method of Croon, we need to extract these two matrices. The observed variance-covariance matrix $\mathbf{S}$ can easily be calculated based on the data. The model implied matrix is not estimated by the method of Croon. However, $\hat{\mathbf{\Sigma}}$ can be approximated by using equation (4.4) from SEM. Since the method of Croon does not estimate the parameters by mimimizing the fit function, we do not have the variance-covariance matrix of the latent variables $\mathbf{\Sigma}_\eta$. However, we do have $\hat{\mathbf{\Sigma}}_\eta$, the matrix with the estimated variances and covariances of the latent variables, based

on the factor scores. Also, the method of Croon does not estimate $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ simultaneously. It estimates $\mathbf{\Lambda}_i$ and $\mathbf{\Theta}_i$ per variable in the structural model. Based on these separate $\mathbf{\Lambda}_i$ and $\mathbf{\Theta}_i$ matrices, a single joint $\mathbf{\Lambda}_c$ and $\mathbf{\Theta}_c$ matrix can easily be constructed by stacking the separate matrices into one big matrix.

Now we can calculate an approximate model implied variance-covariance matrix $\hat{\mathbf{\Sigma}}_a$:

$$\hat{\mathbf{\Sigma}}_a = \mathbf{\Lambda}_c \hat{\mathbf{\Sigma}}_\eta \mathbf{\Lambda}_c^{'} + \mathbf{\Theta}_c. \tag{4.12}$$

Based on $\hat{\mathbf{\Sigma}}_a$ and equation (4.5), an approximate F_{ML_a}-value can be estimated:

$$\mathrm{F}_{ML_a} = tr(\mathbf{S}\hat{\mathbf{\Sigma}}_a^{-1}) + log|\hat{\mathbf{\Sigma}}_a| - log|\mathbf{S}| - k. \tag{4.13}$$

Now, four approximate fit indices can be calculated for the method of Croon:

$$\chi_a^2 = n\mathrm{F}_{ML_a}, \tag{4.14}$$

$$RMSEA_a = \sqrt{max(\frac{\chi_a^2 - \mathrm{df}}{\mathrm{df}(n-1)}, 0)}, \tag{4.15}$$

$$SRMR_a = \sqrt{\frac{\sum_{i=1}\sum_{j=1}[(s_{ij} - \hat{\sigma}_{a_{ij}})/(s_i s_j)]^2}{k(k+1)/2}}, \tag{4.16}$$

$$CFI_a = 1 - \frac{max[(\chi_a^2 - \mathrm{df}_a), 0]}{max[(\chi_0^2 - \mathrm{df}_0), 0]}, \tag{4.17}$$

with the degrees of freedom df_a and the baseline model remaining the same as in SEM.

As in MLE, $\mathbf{S}$ has to be positive-definite to calculate F_{ML_a}, implying that the sample size needs to be at least as large as the number of observed variables to be able to estimate the χ_a^2, the RMSEA or the CFI. However, if we use casewise likelihood, we could compute the likelihood of the saturated model (LL_S) and the model of in-

terest (LL_M). The χ^2 can then be calculated by $-2(LL_M - LL_S)$. The SRMR can be estimated even when the sample size is smaller than the number of observed variables.

Local fit indices

The advantage of using a step-wise method, such as the method of Croon is that we can also evaluate the local fit indices of the measurement models of each latent variable and of the structural model if it is non-saturated. For each factor model that is fitted, the χ^2-value and the other fit indices described above can be calculated. The local fit of the structural model can only be determined if the model is non-saturated. When the model is saturated, the local fit cannot be determined, while the global fit indices are mainly reflections of the local fit indices of the measurement models. In case of misfit, the local fit indices can help us to determine where the problem is situated in the model, while global fit indices only tell us there is a problem. Therefore, we recommend to first look at the global fit of the model. If the global fit is not satisfactory, the local fit indices can be used to localize the problem.

4.3.3 Model comparison test

In SEM, one would typically use a LRT or equivalently a 'difference in χ^2-test' to compare the full model to a restricted model. The only condition is that the restricted model is nested in the full model. This includes models containing restrictions such as equality constraints.

Hereafter, we propose a 'difference in χ^2-test' for factor score regression. FSR does not give us a direct estimator for χ^2. However, in section 3.2.1 of this paper, we introduced χ^2_a for the full model. We can use the same rationale to determine a χ^2_{ar} for the restricted model. However, we need to adjust some of the matrices used.

First, the regression coefficients from the restricted model ($\mathbf{B_r}$) need to be estimated using the method of Croon. Note that only the structural model needs to be respecified and estimated. There is no

need to re-estimate the measurement models, unless there are extra restrictions set on the measurement models. Then, we can calculate the restricted approximated model implied variance-covariance matrix $\hat{\mathbf{\Sigma}}_{ar}$ as follows:

$$\hat{\mathbf{\Sigma}}_{ar} = \mathbf{\Lambda}_c \hat{\mathbf{\Sigma}}_{\boldsymbol{\eta r}} \mathbf{\Lambda}_c^{'} + \mathbf{\Theta}_c, \tag{4.18}$$

with

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{\eta r}} = (\mathbf{I} - \mathbf{B_r})^{-1} \mathbf{\Psi}_{\boldsymbol{r}} (\mathbf{I} - \mathbf{B_r})^{-1^{'}}. \tag{4.19}$$

Using $\hat{\mathbf{\Sigma}}_{ar}$ in formula (4.13) instead of $\hat{\mathbf{\Sigma}}_a$, results in the restricted fit function F_{MLr}, which can be used to calculate χ^2_{ar}:

$$\chi^2_{ar} = nF_{MLr} \tag{4.20}$$

The test statistic for the difference in χ^2_a between the full and the reduced model can be calculated as follows:

$$\begin{aligned}
\chi^2_{ar} - \chi^2_a &= nF_{MLr} - nF_{ML} \\
&= n[F_{MLr} - F_{ML}] \\
&= n[(tr(\mathbf{S}\hat{\mathbf{\Sigma}}_{\boldsymbol{ar}}^{-1}) + log|\hat{\mathbf{\Sigma}}_{\boldsymbol{ar}}| - log|\mathbf{S}| - k) \\
&\qquad -(tr(\mathbf{S}\hat{\mathbf{\Sigma}}_a^{-1}) + log|\hat{\mathbf{\Sigma}}_a| - log|\mathbf{S}| - k)] \\
&= n[tr(\mathbf{S}\hat{\mathbf{\Sigma}}_{\boldsymbol{ar}}^{-1}) + log|\hat{\mathbf{\Sigma}}_{\boldsymbol{ar}}| - log|\mathbf{S}| - k \\
&\qquad -tr(\mathbf{S}\hat{\mathbf{\Sigma}}_a^{-1}) - log|\hat{\mathbf{\Sigma}}_a| + log|\mathbf{S}| + k] \\
&= n[tr(\mathbf{S}\hat{\mathbf{\Sigma}}_{\boldsymbol{ar}}^{-1}) + log|\hat{\mathbf{\Sigma}}_{\boldsymbol{ar}}| \\
&\qquad -tr(\mathbf{S}\hat{\mathbf{\Sigma}}_a^{-1}) - log|\hat{\mathbf{\Sigma}}_a|]
\end{aligned} \tag{4.21}$$

Note that $log|\mathbf{S}|$ cancels out, which implies that the requirement that $\mathbf{S}$ is positive definite no longer holds. As a consequence the difference between χ^2_a and χ^2_{ar} can, in principle, be estimated when the sample size is smaller than the number of observed variables. The difference $\chi^2_{ar} - \chi^2_a$ follows a distribution that closely approximates a χ^2-distribution with the degrees of freedom equal to the

number of regression coefficients that are restricted to 0. Consequently, we use this distribution to test if the difference score is significantly different from 0. A significant test yields evidence that the variable has significant effect and that the full model should be preferred.

One of the advantages of using a model comparison test, is that it can also be used for categorical variables. It is widely known that a categorical variable with d classes can be recoded into a set of $d-1$ dummy (0/1) variables. The covariances between factor scores and dummy-coded predictor variables can be corrected by using Croon's formulas in the same manner as continuous variables. The main issue with using these dummy variables is that an omnibus test is needed to test the significance of the categorical variable, if d is larger than 2. Using an omnibus test, we can evaluate if all regression coefficients of all $d-1$ dummy variables are 0. This can be done by using the 'difference in χ^2-test' that we have proposed, but now the regression coefficients of all dummy variable are set to 0 in the reduced model. The test statistic then follows a χ^2-distributions with $d-1$ degrees of freedom.

4.4 Simulation study

In this section, we conduct a simulation study to compare the new fit indices of FSR with their counterparts obtained by SEM using MLE and study the performance of our newly proposed model comparison test in terms of type I error rate and power. We expect that our fit indices will closely resemble the fit indices obtained by using MLE. Further, we evaluate the performance of the method of Croon for models with categorical predictors in terms of bias of the regression coefficients.

4.4.1 Data generating mechanism

The model in Figure (4.1) was used to simulate the data in R (R Core Team, 2016). First the four predictor variables were generated

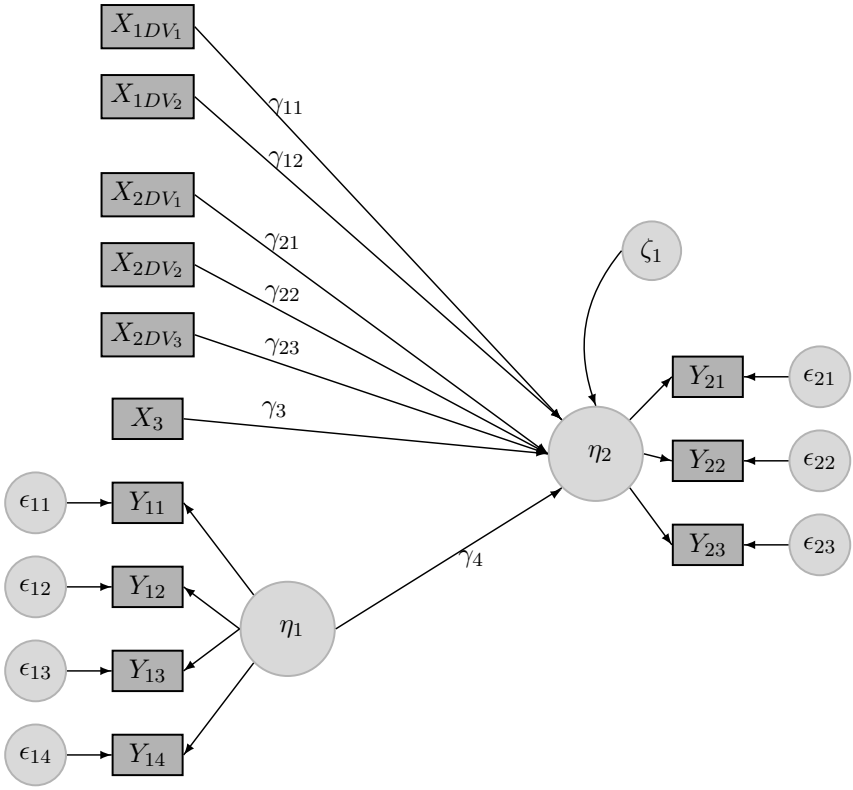


Figure 4.1 The ground truth model for the simulation study.

from a multivariate normal distribution:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \eta_1 \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & .50 & .10 & .30 \\ .50 & 1.00 & .25 & .25 \\ .10 & .25 & 1.00 & .50 \\ .30 & .25 & .50 & 1.00 \end{bmatrix} \right)$$

The variable correlations were chosen to range from strong (.5), moderate (.25-.30) and weak (.1) correlations. Variable X_1 and X_2 are considered as categorical with three and four categories respectively. Variables X_3 and η_1 are considered as continuous, where η_1 is a latent variable measured by four indicators. To create the categories for variables X_1 and X_2, we select two (three for X_2) cut-off points. For X_1 these cut-off points are quantile 25 and 65, which implies an unbalanced design with on average 25% in the first category, 40% in the second and 35% in the third. The cut-off points for the second categorical variable are the $35^{th}, 70^{th}$ and 85^{th} quantile.
Using the structural equation

$$\eta_1 = \gamma_{11} X_{1DV_1} + \gamma_{12} X_{1DV_2} + \gamma_{21} X_{2DV_1} + \gamma_{22} X_{2DV_2} + \gamma_{23} X_{2DV_3} + \gamma_3 X_3 + \gamma_4 \eta_1 + \zeta_1$$

the true latent scores on η_1 were generated. The following parameter settings were used: $\gamma_{11} = .3, \gamma_{12} = .7$ and $\gamma_4 = .5$. To control the type I error rate we assume no effect of the second categorical variable and the first continuous variable. Therefore the parameters which belong to the dummy- variables used to code this categorical variable and the first categorical variables X_3 were set to 0 (i.e. $\gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_3 = 0$)
The indicator variables for the latent variables η_1 and η_2 were generated using the measurement models:

$$\mathbf{y} = \mathbf{\Lambda \eta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ was generated from a multivariate normal distribution with means $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Theta}$. $\mathbf{\Theta}$ is a diagonal matrix

with the residual variances θ_{li} on the diagonal. The index 'i' is used to refer to the items, while index 'l' refers to the latent variables. The θ_{li}-values are calculated as follows:

$$\theta_{li} = \frac{\text{var}(\eta_l)(1 - CD_{li})}{CD_{li}}, \tag{4.22}$$

with CD_{li} the coefficients of determination, which give us an indication of the reliability of the factor loadings. All CD_{li} are set to be equal and are varied between 0.5, 0.7 and 0.9. The sample size is varied between 50,100, 200, 500 and 1000, resulting in 15 different conditions in this simulation study. For each condition we generated 2000 datasets.

4.4.2 Analysis

Each of the 30 000 datasets were analyzed with three methods: the uncorrected FSR, FSR with Croon's corrections, and SEM using MLE. The significance level (α) was set to .05. For each data set, five models were fitted for each method. To compare the power of the three methods, a reduced model without the first categorical predictor (i.e. without the set of all X_1 dummy-coded variables, further referred to as Red_{X_1}) and a reduced model without the second continuous variable are fitted and compared to the full model with all exogenous variables. Further, we study the Type I error rate by fitting a model without the second categorical variable ((i.e. without the set of all X_2 dummy-coded variables, further referred to as Red_{X_2}) and a model without the first continuous variable (further referred to as Red_{X_3}). Since the regression coefficients for these variables are set to 0 in the data generating mechanism, the full model should be preferred in only five percent of the cases.
For conducting the uncorrected FSR, a factor analysis was performed for the latent variables with the cfa-function of the lavaan-package in R (Rosseel, 2012). Then factor scores were calculated using the regression predictor. Next the predicted factor scores are used in a linear regression using the lm-function to regress the pre-

dicted factor scores of Y on the exogenous variables. For FSR with Croon's corrections, the same cfa-function and factor scores were used, but Croon's corrections were applied to address the bias in the variance-covariance matrix of the factor scores. Estimators for the regression coefficients were obtained directly from the corrected variance-covariance matrix. Finally, a MLE analysis was conducted with the sem-function in lavaan.
For each method, we consider the bias in the regression coefficients, the power and type I-error rate of the model comparison test and the different fit indices. Four fit indices were considered, namely the χ^2-value, RMSEA, SRMR and CFI.

4.4.3 Results

Bias Figure 4.2 shows the results for the estimators of the regression coefficients for each method averaged over 2000 simulations. The findings are in line with the expectations. More specifically, MLE shows no bias while the uncorrected FSR shows bias. This bias is independent of sample size, but is influenced by the coefficient of determination (CD). The FSR with Croon's corrections corrects for the bias and thus shows results in line with MLE.
χ^2-value Table 4.1 shows the χ^2-values and χ^2_a-values for each model. We can see that the newly proposed χ^2_a-value approximates the original χ^2 calculated by MLE rather well. The deviations are in the same direction and have the same magnitude for all models. Consequentially, the test statistic to test if the categorical variable is significant, i.e. the difference in χ^2_a-values, is almost identical as the observed χ^2- test statistic in SEM, i.e. the difference in χ^2-values. Based on these results, it can be expected that the power of both methods will be very similar.
Fit indices The results of the other three fit indices for the full model are shown in Table 4.3 (Results for other models are similar, but not shown). These results show that the approximated fit indices obtained by the FSR approach are similar to the fit indices estimated by MLE.

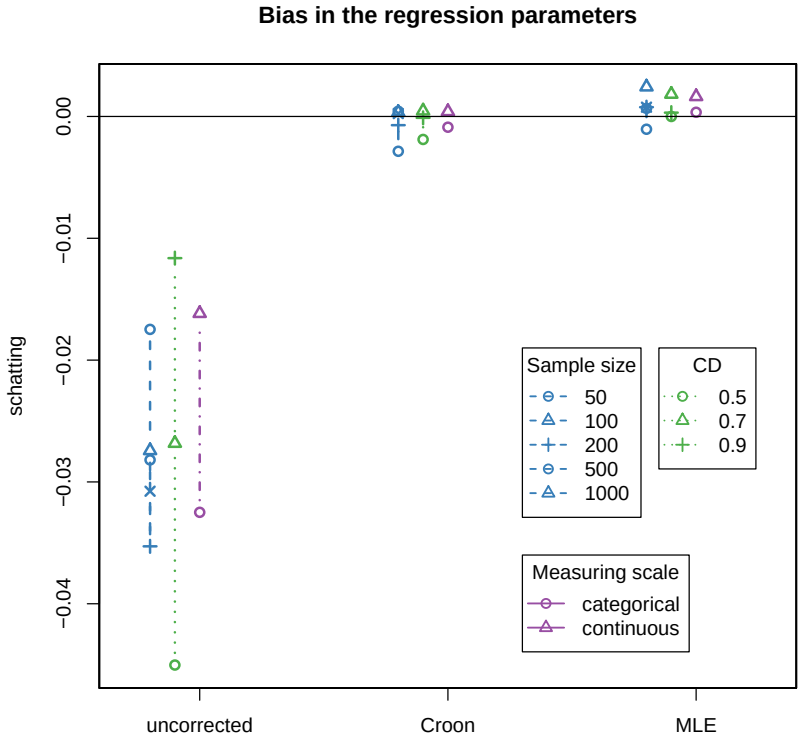


Figure 4.2 The influence of sample size, coefficient of determination and measurement scale on the bias in the regression coefficients. Uncorrected denotes the latent approach without bias correction, Croon denotes the latent approach with Croon's corrections and MLE denotes the results when data is analyzed with maximimum likelihood estimation. The black horizontal line denotes a bias of 0.

Table 4.1 Results for χ^2_a and χ^2 for the full model (Full), the reduced model without the first (second) categorical variable, Red_{X_1} (Red_{X_2}) and the reduced model without the first (second) continuous variable, Red_{X_3} (Red_{η_1}).

		Full		Red_{X_1}		Red_{X_2}		Red_{X_3}		Red_{η_1}	
CD	N	χ^2_a	χ^2	χ^2_a	χ^2	χ^2_a	χ^2	χ^2_a	χ^2	χ^2_a	χ^2
.5	50	51.25	50.62	58.02	57.47	54.74	54.21	52.50	41.95	54.34	53.85
.5	100	47.13	46.25	55.90	55.17	50.41	49.60	48.21	47.34	51.72	51.03
.5	200	45.02	44.37	59.05	58.46	48.08	47.45	46.09	45.44	53.56	53.08
.5	500	44.62	43.96	75.57	74.93	47.69	47.02	45.65	44.96	61.91	61.37
.5	1000	43.71	43.08	106.69	105.97	46.79	46.15	44.72	44.06	76.05	75.53
.7	50	51.01	50.42	56.62	56.18	54.55	54.08	52.19	51.63	55.32	54.89
.7	100	47.48	47.04	55.72	55.36	50.74	50.33	48.56	48.12	52.90	52.54
.7	200	44.93	44.63	61.06	60.82	48.17	47.89	46.00	45.70	55.52	55.29
.7	500	44.28	43.96	84.13	83.85	47.35	47.03	45.35	45.01	66.27	66.01
.7	1000	43.40	43.10	117.08	116.77	46.40	46.10	44.44	44.13	86.97	86.73
.9	50	50.86	50.75	57.27	57.20	54.39	54.30	52.04	51.92	55.72	55.66
.9	100	47.24	47.13	56.28	56.19	50.50	50.39	48.32	48.22	53.72	53.63
.9	200	44.83	44.75	63.64	63.58	47.88	47.80	45.88	45.80	56.61	56.55
.9	500	43.79	43.72	86.06	86.01	46.88	46.81	44.78	44.71	73.34	73.29
.9	1000	43.21	43.15	124.00	123.94	46.16	46.10	44.24	44.18	97.87	97.80

Table 4.2 Results for the test statistics obtained by FSR and SEM using MLE averaged over 2000 simulations. Red_{X_1} (Red_{X_2}) denotes differences in χ^2_a and χ^2 for the reduced model without the first (second) categorical variable in comparison with the full model. Red_{X_3} (Red_{η_1}) denotes the differences between the reduced model without the first (second) continuous variable and the full model.

		Red_{X_1}		Red_{X_2}		Red_{X_3}		Red_{η_1}	
CD	N	$Dif_{\chi^2_a}$	Dif_{χ^2}	$Dif_{\chi^2_a}$	Dif_{χ^2}	$Dif_{\chi^2_a}$	Dif_{χ^2}	$Dif_{\chi^2_a}$	Dif_{χ^2}
.5	50	6.77	6.86	3.49	3.57	1.24	1.26	3.08	3.21
.5	100	8.77	8.95	3.28	3.35	1.09	1.09	4.59	4.79
.5	200	14.02	14.09	3.06	3.08	1.07	1.07	8.54	8.72
.5	500	30.94	30.94	3.06	3.06	1.02	1.01	17.29	17.41
.5	1000	62.98	62.89	3.08	3.07	1.01	0.99	32.34	32.45
.7	50	5.61	5.72	3.54	3.61	1.18	1.20	4.31	4.46
.7	100	8.24	8.32	3.26	3.29	1.08	1.08	5.41	5.50
.7	200	16.13	16.19	3.25	3.26	1.08	1.08	10.59	10.66
.7	500	39.85	39.89	3.07	3.08	1.07	1.06	21.99	22.06
.7	1000	73.68	73.67	3.00	3.00	1.04	1.04	43.58	43.64
.9	50	6.41	6.43	3.52	3.54	1.17	1.18	4.85	4.89
.9	100	9.04	9.06	3.25	3.26	1.08	1.08	6.47	6.50
.9	200	18.81	18.83	3.05	3.05	1.05	1.05	11.78	11.80
.9	500	42.27	42.29	3.09	3.09	0.99	0.99	29.55	29.57
.9	1000	80.79	80.79	2.95	2.95	1.03	1.03	54.66	54.66

Power Table 4.4 shows the results of the model comparisons between the full model and the four reduced models. As mentioned before, the power can only be compared for models Red_{X_1} and Red_{η_1}. Since the full model should be preferred, the number of p-values smaller than .05 gives an indication of the power of each method. As expected, we see that there is almost no difference between MLE and the method of Croon. The power is similar for both methods.

Type I error rate The results in Table 4.4 shows that MLE and the method of Croon perform similar in terms of type I error rate. They both approximate the true level of 0.05, when the sample size is at least 100. When the sample size drops below 100, the type I error rate is inflated. The inflation is worse for MLE than for the method of Croon and is also worse for the categorical variables than for continuous variables.

4.5 Illustration

We use data from the âĂŸMonitoring ICT integration in Flemish Education 2012âĂŹ (MICTIVO2) to illustrate the method of Croon, with the newly developed fit indices and model comparison tests. This study assesses the impact of ICT at all levels of formal education Goeman, Elen, Pynoo, and van Braak (2015). The latent constructs in the study were all measured using several indicator variables, using Likert scales, ranging from 5-point to 7-point scales.

Data was collected by using several surveys, from 4887 students, 2585 teachers and 723 principals from 729 schools. Only pupils from primary education and their principals were used in this paper. The pupils were only selected if their respective principal also participated in the study. This resulted in 2033 pupils and 47 principals.

Table 4.3 Fit indices for the Full model estimated by FSR and MLE over 2000 simulations.

		CFI		SRMR		RMSEA	
CD	N	Croon	MLE	Croon	MLE	Croon	MLE
.5	50	.964	.966	.058	.058	.055	.052
.5	100	.980	.983	.050	.049	.028	.026
.5	200	.993	.994	.035	.034	.016	.015
.5	500	.998	.998	.022	.022	.010	.009
.5	1000	.999	.999	.016	.015	.006	.006
.7	50	.973	.974	.051	.050	.053	.050
.7	100	.990	.990	.037	.037	.029	.028
.7	200	.996	.996	.027	.027	.016	.016
.7	500	.998	.999	.017	.017	.009	.009
.7	1000	.999	.999	.012	.012	.006	.006
.9	50	.983	.983	.032	.032	.053	.052
.9	100	.995	.995	.021	.021	.029	.028
.9	200	.998	.998	.016	.016	.016	.016
.9	500	.999	.999	.010	.010	.009	.009
.9	1000	1	1	.007	.007	.006	.006

Table 4.4 Type I error rate and power over 2000 simulations for uncorrected latent method (Lat), FSR with Croon's corrections (CR) and SEM using MLE with signficance level $\alpha = .05$.

		Power						Type I error rate					
		X_1			η_1			X_2			X_3		
CD	N	Lat	CR	MLE	Lat	CR	MLE	Lat	CR	MLE	Lat	CR	MLE
.5	50	.393	.474	.480	.236	.291	.304	.052	.084	.090	.054	.080	.083
.5	100	.594	.644	.659	.444	.476	.487	.049	.063	.071	.049	.055	.055
.5	200	.892	.891	.892	.788	.798	.804	.038	.045	.047	.063	.062	.062
.5	500	.999	.999	.999	.979	.979	.980	.045	.050	.050	.090	.053	.049
.5	1000	1	1	1	1	1	1	.053	.055	.053	.140	.047	.046
.7	50	.291	.377	.381	.372	.436	.456	.051	.082	.094	.047	.070	.072
.7	100	.530	.600	.601	.513	.550	.556	.054	.067	.069	.045	.058	.059
.7	200	.931	.937	.939	.870	.879	.884	.058	.065	.066	.060	.059	.058
.7	500	1	1	1	.998	.998	.997	.052	.057	.055	.077	.056	.056
.7	1000	1	1	1	1	1	1	.048	.049	.050	.086	.052	.052
.9	50	.358	.449	.451	.419	.493	.495	.055	.087	.088	.045	.070	.070
.9	100	.604	.656	.657	610	.641	.647	.051	.068	.069	.051	.060	.060
.9	200	.969	.974	.974	.901	.908	.907	.051	.061	.061	.068	.061	.061
.9	500	1	1	1	1	1	1	.051	.053	.053	.043	.044	.044
.9	1000	1	1	1	1	1	1	.047	.048	.048	.054	.050	.050

4.5.1 Model 1

For the first model, we only use the data from the pupils. We consider two latent variables, ICT-skills of the pupils and ICT-use by the pupils. Both variables were measured using 6 indicator items. We are interested in the relationship between both variables, while controlling for the variables gender and language (mother tongue). Both covariates are considered as categorical. Gender consists of two categories, with female used as the reference level. Language consists of three categories, namely speaking a Dutch dialect, Dutch or another language. Two dummy variables were created. Language1 represents speaking a dialect, while language2 represents speaking Dutch. Speaking another language is the reference level. The model is depicted in Figure 4.3 and the results can be found in table 4.5. It can be seen that the Croon estimates are close to the MLE-estimates for the regression parameters of both the continuous and categorical variables. Even more importantly, the p-values and fit indices are also close together and result in the same conclusions.

4.5.2 Model 2

For the second model, we use the data from the 47 principals. We consider four latent variables, namely ICT-use of the teachers, the pedagogical ICT-competencies of the teachers, the quality of the ICT-policy of the school and the ICT-professionalization of the teachers. All variables are assessed by the principals. All variables were measured using several indicator items. In total, there are 49 indicators. We are interested in the influence of the competencies, ICT-policy and professionalization on the ICT-use of the teachers. The model is depicted in Figure 4.4. Since there are more observed variables (p=49), than observations (N=47), standard MLE based on Eq.(4.5) could not estimate the regression parameters. The method of Croon was able to estimate the parameters of interest, as can be seen in Table 4.6. Further, the test statistic for model comparisons and an estimate for SRMR can be obtained using FSR. An estimator for χ^2_a for the full or a reduced model is not

Table 4.5 The estimated parameters and their standard errors for the first illustration, both for the Croon method and MLE. The p-values that are estimated using a model comparison test are indicated with an asterisk. Note that the parameter estimates and standard errors of the measurement models result from the separate factor analyses for the Croon method.

		Croon			MLE	
Parameter	Est.	(SE)	p-value	Est.	(SE)	p-value
Parameter estimates						
Measurement models						
ICT-skills						
λ_1	1			1		
λ_2	1.020	0.026	0.000	1.026	0.027	0.000
λ_3	0.962	0.030	0.000	0.973	0.030	0.000
λ_4	0.900	0.030	0.000	0.908	0.030	0.000
λ_5	0.606	0.023	0.000	0.612	0.023	0.000
λ_6	0.562	0.022	0.000	0.566	0.021	0.000
ICT-use						
λ_7	1			1		
λ_8	0.693	0.040	0.000	0.725	0.040	0.000
λ_9	0.831	0.048	0.000	0.862	0.047	0.000
λ_{10}	0.893	0.046	0.000	0.906	0.050	0.000
λ_{11}	0.549	0.039	0.000	0.570	0.038	0.000
λ_{12}	0.747	0.061	0.000	0.833	0.061	0.000
Regressions						
ICT-use	0.405		0.000*	0.422	0.035	0.000
gender	0.139		0.001*	0.141	0.044	0.001
language1	0.187		0.014*	0.183	0.076	0.015
language2	0.080		0.200*	0.082	0.062	0.191
Omnibus test language						
language			0.042*			0.049*
fit indices						
χ^2	491.522		0.038	488.435		0.039
RMSEA	0.049			0.049		
SRMR	0.038			0.036		
TLI	0.934			0.935		

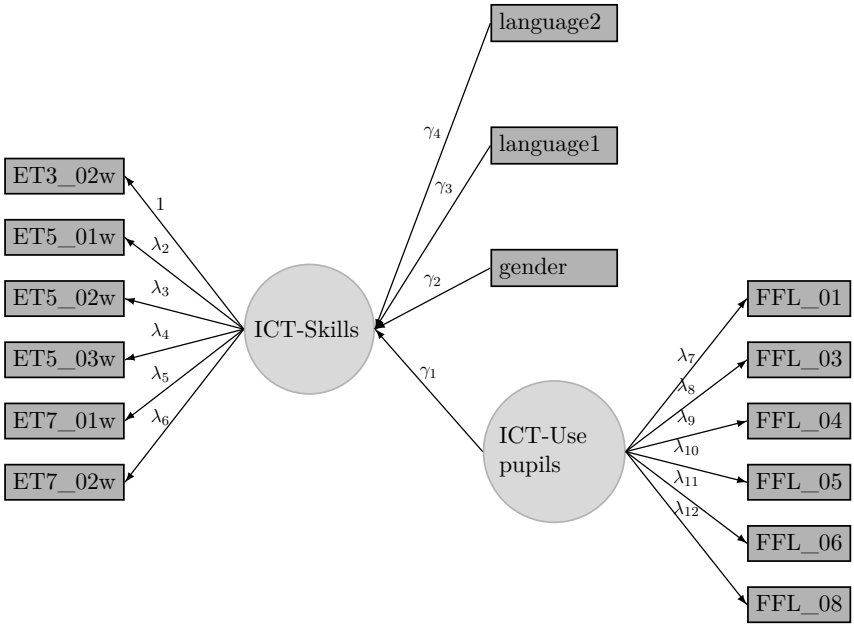


Figure 4.3 The first model used in the illustration.

possible since **S** is not positive definite.

4.6 Discussion

While it has been shown that FSR without correction yields biased estimators for the regression coefficients, this method remains often used by applied researchers. A correction for this bias was developed by (Croon, 2002). Devlieger and Rosseel (2017) and Lu et al. (2011) have shown for continuous variables that FSR with Croon's corrections performs at least as well as MLE. In several settings FSR with Croon's corrections even outperforms MLE, such as small to moderate sample sizes, models with misspecifications and multilevel settings with a small number of clusters.

In this paper, we introduced two extensions to this method. First, we introduced fit indices for FSR that approximate their counterparts that are typically reported in SEM using MLE. Consequently,

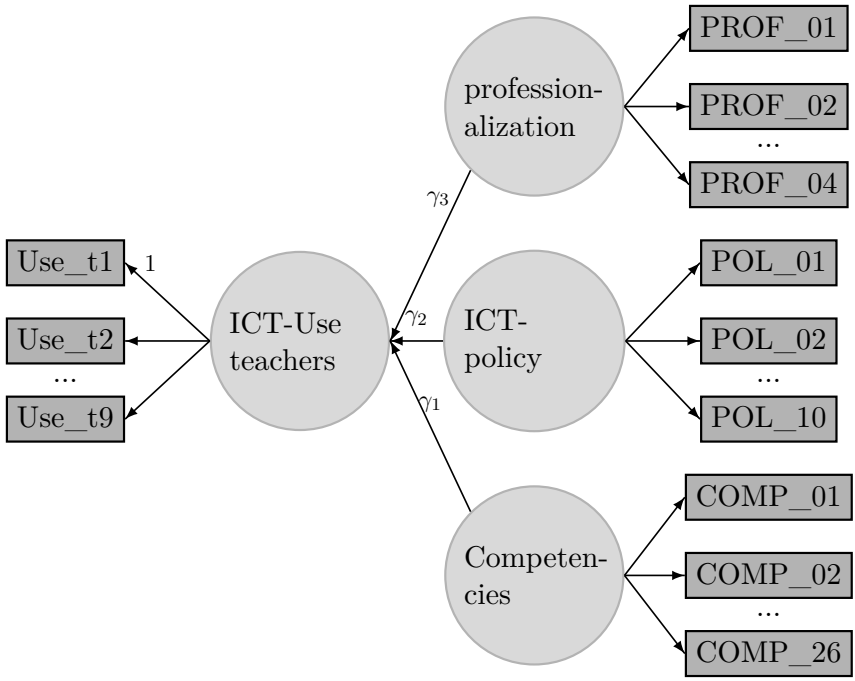


Figure 4.4 The second model used in the illustration.

Table 4.6 The estimated regression parameters and their p-values for the second illustration, for the Croon method. The p-values that are estimated using a model comparison test are indicated with an asterisk.

Parameter	Est.	p-value
Parameter estimates		
Regressions		
Pedagogical ICT-competencies	0.639	0.632*
ICT-policy	-0.010	0.952*
Professionalisation	0.228	0.650*
fit indices		
SRMR	0.185	

the global fit of the model can now be assessed. Note that the local fit of the different parts of the model can also be evaluated when using FSR. In a simulation study, we have shown that the newly developed fit indices are very similar to their counterparts estimated using MLE.

Secondly, we proposed a model comparison test based on the new fit index χ^2_a for FSR, which can be used to compare two nested models. We have shown that this test for FSR performs well in terms of type I error rate and power. The model comparison test can also be used to draw inference for a single continuous predictor, without the need for a standard error. Derivations for the standard errors are not yet available in the current literature about FSR. However, an analytical solution is under development (Rosseel, 2019).

A major advantage of the Croon-based model comparison over MLE is that the method of Croon can also estimate models where the sample size is smaller than the number of observed variables. However, since we approximate the *"classic"* χ^2 from MLE, we also inherit the problems that come with it. More specifically, it is widely known that the chi-square test leads to inflated type I error rates when the sample size is small (Nevitt & Hancock, 2004). Our simulation study did indeed show an inflated type I error rate when the sample size is small. In the literature, several corrections have been proposed to improve the performance of the χ^2- test statistic, such as the Bartlett correction (Bartlett, 1937, 1954; Savalei, 2010) and the Swain correction (Swain, 1975). More research is needed to determine if these corrections could also be applied to the approximated χ^2_a for FSR.

To evaluate the fit of a SEM-model, an alternative to the χ^2- test has recently been proposed by Maydeu-Olivares (2017). He proposes to use a significance test that is based on the SRMR (Maydeu-Olivares, 2017; Maydeu-Olivares, Shi, & Rosseel, 2018). Since we also proposed an approximate SRMR, this method could possibly also be applied to FSR. More research is needed to evaluate the performance of this SRMR significance test in FSR. Finally, the

performance of all proposed fit indices should be further studied to determine how they behave in case of model misspecification, low sample sizes, non-normal data or missing data.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, *103*(3), 411–423. doi: 10.1037/0033-2909.103.3.411

Bakk, Z., & Kuha, J. (2017). Two-Step Estimation of Models Between Latent Classes and External Variables. *Psychometrika*, 1–22. doi: 10.1007/s11336-017-9592-7

Bartlett, M. (1937). Properties of Sufficiency and Statistical Tests. *Proc. R. Soc. Lond. A*, *160*(901), 268–282.

Bartlett, M. (1954). A Note on the Multiplying Factors for Various $\chi 2$ Approximations. *Journal of the Royal Statistical Society Series B (Methodological)*, *16*(2), 296–298.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*, 78–117. doi: 10.1177/0049124187016001004

Bollen, K. (1989). *Structural equations with latent variabels*. New York: NY: Wiley.

Bollen, K. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109–121. doi: 10.1007/BF02296961

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi: 10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology*, *13*, 31–38. doi: 10.1027/1614-2241/a000130

Devlieger, I., & Rosseel, Y. (2019). *Multilevel factor score regression (Under review).*

Gagne, P., & Hancock, G. R. (2006). Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. *Multivariate Behavioral Research*, *41*(1), 55–64. doi: 10.1207/s15327906mbr4101

Goeman, K., Elen, J., Pynoo, B., & van Braak, J. (2015). Time for action! ICT Integration in Formal Education: Key Findings from a Region-wide Follow-up Monitor. *TechTrends*, *59*(5), 40–50. doi: 10.1007/s11528-015-0890-6

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. doi: 10.1080/10705519909540118

Huang, P. H., Chen, H., & Weng, L. J. (2017). A Penalized Likelihood Method for Structural Equation Modeling. *Psychometrika*, *82*(2), 329–354. doi: 10.1007/s11336-017-9566-9

Jacobucci, R., Grimm, K. J., Brandmaier, A. M., Serang, S., & Kievit, R. A. (2018). *regsem : Regularized Structural Equation Modeling.* Retrieved from `https://cran.r-project.org/package=MIIVsem`

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social science.s* (pp. 85–112). New York: Seminar Press. doi: 10.1002/j.2333-8504.1970.tb00783.x

Kaplan, D. (1955). Path Analysis: Modeling Systems of Structural Equations Among Observed Variables. In *Structural equation modeling: Foundations and extensions* (pp. 13–39). Sage Publications.

Lastovicka, J. L., & Thamodaran, K. (1991). Common factor score estimates in multiple regression problems. *Journal of Marketing Research*, *28*(1), 105–112. doi: 10.2307/3172730

Lewis, J. B. (2005, jul). Estimating regression models in which the dependent variable Is based on estimates. *Political Analysis*, *13*(4), 345–364. doi: 10.1093/pan/mpi026

Lu, I. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011, sep). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, *28*(3), 258–268. doi: 10.1016/j.ijresmar.2011.03.006

Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models : A comparison with regression based on IRT scores. *Structural equation modeling: A multidisciplinary Journal*, *12:2*(January 2014), 263–277. doi: 10.1207/s15328007sem1202

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling*, *5511*(11:3), 320–341. doi: 10.1207/s15328007sem1103

Maydeu-Olivares, A. (2017). Assessing the Size of Model Misfit in Structural Equation Models. *Psychometrika*, *82*(3), 533–558. doi: 10.1007/s11336-016-9552-7

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing Fit in Structural Equation Models : A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit Assessing Fit in Structural Equation Models : A Monte-Carlo Evaluation of RMSEA Versus SRMR Con fi dence Intervals and. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 389–402. Retrieved from `https://doi.org/10.1080/10705511.2017.1389611` doi: 10.1080/10705511.2017.1389611

McNeish, D. (2016). On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling*, *23*(5), 750–773. doi: 10.1080/10705511.2016.1186549

Nevitt, J., & Hancock, G. R. (2004). Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modeling Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modeling. *Multivariate Behavioral Research*, *39*(2), 439–478. doi: 10.1207/S15327906MBR3903

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical computing. Retrieved from `https://www.r-project.org/`

Rosseel, Y. (2012). lavaan : An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Rosseel, Y. (2019). *Analytic standard errors for factor score regression with Croon's corrections. Manuscript in preparation.*

Savalei, V. (2010). Small Sample Statistics for Incomplete Nonnormal Data : Extensions of Complete Data Formulae and a Monte Carlo Comparison Small Sample Statistics for Incomplete Nonnormal Data : Extensions of Complete Data Formulae and a Monte Carlo Comparison. *Structural Equation Modeling*, *17*(2), 241–264. doi: 10.1080/10705511003659375

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian Estimation and Testing of Structural Equation Models. *Psychometrika*, *64*(June), 37–52.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74. doi: 10.1002/0470010940

Schumacker, R., & Lomax, R. (1996). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shevlin, M., Miles, J. N. V., & Bunting, B. P. (1997). Summated rating scales. A Monte Carlo investigation of the effects of reliability and collinearity in regression models. *Personality and Individual Differences*, *23*(4), 665–676. doi: 10.1016/S0191-8869(97)00088-3

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. doi: 10.1007/BF02296196

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Unpublished doctoral dissertation). University of Adelaide, Department of Statistics.

5 Multilevel factor score regression

Abstract. Multilevel SEM is an increasingly popular technique to analyze data that are both hierarchical and contain latent variables. The parameters are usually jointly estimated using a maximum likelihood estimator (MLE). This has the disadvantage that a large sample size is needed and misspecifications in one part of the model may influence the whole model. We propose an alternative stepwise estimation method, which is an extension of the Croon method for factor score regression (Croon, 2002). In this paper, we extend this method to the multilevel setting. A simulation study was used to compare this new estimation method to the standard MLE. The Croon method outperformed MLE with regard to convergence rate, bias, MSE, and coverage, in particular when models contained a structural misspecification. In conclusion, the Croon method seems to be a promising alternative to MLE.

This chapter has been published as Devlieger, & Rosseel, Y. (2019). Multilevel factor score regression, *Multivariate Behavioral Esearch*, doi: 10.1080/00273171.2019.1661817.

5.1 Introduction

In the social and behavioral sciences, data often have a hierarchical structure. A common example is the clustering of pupils within schools. Many of the variables of interest are also latent, which means they cannot be measured directly. Examples are reading ability, intelligence, motivation, etc. These latent variables are usually measured using observed indicators. Although these indicators can be continuous or categorical, this paper will only focus on latent variables with continuous indicators. Multilevel Structural Equation Modeling is often used to study the relationships between latent variables with a hierarchical structure. Over the years, many approaches to multilevel SEM have been developed (e.g. Goldstein & McDonald, 1988; Lee, 1990; McDonald & Goldstein, 1989; B. Muthén, 1989, 1994; B. Muthén & Asparouhov, 2009; Rabe-hesketh, Skrondal, & Andrew, 2004) resulting in several modeling frameworks, such as the within-between framework (Goldstein & McDonald, 1988) and the generalized linear latent and mixed (GLLAMM)-framework. Regardless of the framework, several estimators can be used. For continuous data, maximum likelihood estimation (MLE) is most commonly used. MLE has two main drawbacks. First, misspecifications in one part of the model could influence the whole model (Bollen, 1996). For example, misspecifications in the measurement model may bias the estimates in the structural model. Second, a large sample size is required to obtain unbiased estimators (Schumacker & Lomax, 1996; Valluzzi, Larson, & Miller, 2003).

In the single level context, several alternative methods have been developed to overcome these limitations. Bollen (1996, 2018) developed the instrumental variables approach, which overcomes the misspecification issue. A second alternative is a two-step SEM method, where the measurement models are estimated independently in a first step. In a second step the full SEM-model is estimated with the parameters from the measurement models fixed to the values obtained in the first step (Hunter & Gerbing, 1982; Lance, Cornwell,

& Mulaik, 1988). As a result, misspecifications in the measurement model no longer propagate to the structural model. A third alternative that is often used by applied researchers is "factor score regression". In a first step, the measurement models are estimated and factor scores are computed. Subsequently, these factor scores are used in a linear regression analysis. Factor score regression has one major disadvantage, namely it results in biased estimates of the regression parameters. This is caused by the discrepancy between the (co)variances of the computed factor scores and the (co)variances of the true latent variable scores (Croon, 2002; Dijkstra, 2010).

Several methods were developed that avoid or correct for this bias. When using factor analysis to estimate the measurement model, three main approaches can be used. A first approach was developed by Skrondal and Laake (2001). They avoided the bias by using different predictors to compute the factor scores. For independent latent variables, they used the Regression predictor (Thomson, 1934; Thurstone, 1935) and for dependent variables they used the Bartlett predictor (Bartlett, 1937; Thomson, 1938). While this method resulted in unbiased unstandardised regression parameters, the standardised regression parameters were still biased (Devlieger, Mayer, & Rosseel, 2016). A second approach was developed by Croon (2002). He estimated the (co)variances of the true latent variable scores, and used these to calculate the regression parameters. The method of Croon could be performed using any predictor or estimation method. Hoshino and Bentler (2013) proposed a very similar approach. A disadvantage of their method is that it only works with the Bartlett predictor and relies on weighted least squares (WLS) estimation (Devlieger & Rosseel, 2017). In addition, their method cannot easily be extended to the multilevel setting. When using partial least squares (PLS) to estimate the measurement model, Dijkstra (2010) proposed consistent PLS (PLSc), which uses an algorithm that performs a correction of the correlations between the latent constructs to make results consistent with a factor-model.

Both Lu, Kwan, Thomas, and Cedzynski (2011) and Devlieger et al. (2016) compared the method of Croon, the method of Skrondal and Laake, and MLE to each other using linear regression. Lu et al. (2011) concluded that the Croon method showed relative bias, mean absolute deviation, coverage, and power that is at least comparable to those of the other methods and often much better. Especially in terms of bias, the Croon method outperformed the other methods, when the sample size is small. Devlieger et al. (2016) concluded that the Croon method was the best suitable alternative to MLE, based on the bias, efficiency, mean square error, power, and type I-error rate. Takane and Hwang (2018) compared the performance of the method of Croon, the method of Skrondal and Laake, and PLSc. With regard to the bias for the regression parameters, the method of Skrondal and Laake was outperformed by the other two methods, which gave very similar results. However, the method of Croon clearly outperformed PLSc when it came to estimating the measurement model. Devlieger and Rosseel (2017) showed that the method of Croon indeed outperforms MLE with regard to its two main issues; in terms of bias and convergence rate, it handles small sample sizes and misspecifications better than MLE. Clearly, the method of Croon is a good alternative to SEM in the single level setting. Therefore, we will focus on this method.

The goal of this paper is to extend the method of Croon to the multilevel setting, since the issues in single-level SEM using MLE translate to the multilevel setting. Lüdtke, Marsh, Robitzsch, and Trautwein (2011) stated that the parameters at the between level tend to be biased when only limited information on the between level is available (e.g low intra-class correlations (ICC), small number of clusters, and small number of elements within groups.) Based on a simulation study, Meuleman and Billiet (2009) suggested that at least 60 groups are needed at the between level if one is interested in large effects and at least 100 groups are needed if one is interested in smaller effect sizes. Hox, Maas, and Brinkhuis (2010) also suggested at least 100 groups. Small sample sizes on the between level

not only pose problems with regard to bias, Li and Beretvas (2013) also found serious convergence issues when there were fewer than 80 clusters. Recently, Bayesian approaches have been explored to address these sample size issues (Depaoli & Clifton, 2015; Zitzmann, Lüdtke, Robitzsch, & Marsh, 2016). However, the use of inaccurate priors may have an adverse effect, especially in small samples (Holtmann et al., 2016; van Erp, Mulder, & Oberski, 2018). Holtmann et al. (2016) compared MLE, weighted least squares, mean and variance adjusted (WLSMV), and Bayesian methods for multilevel SEM in small sample sizes. When using continuous indicators, no performance advantages of the Bayesion estimation methods over MLE were found. When using categorical indicators, severe bias was found when strong informative inaccurate priors were used.

Since all parameters are still estimated simultaneously, now across different levels, the misspecification issue also holds. It has been shown that stepwise estimation methods can overcome these issues in the multilevel setting. Goldstein (1987) used multivariate multilevel modeling to estimate the between–group and the within–group covariance matrix. In the second stage, a structural equation model was fitted to each covariance matrix separately. Chou, Bentler, and Pentz (2000), on the other hand, estimated a structural equation model for each cluster separately. The estimates were then used as response values in a between model. Asparouhov and Muthén (2010) implemented a stepwise approach based on plausible values. Zitzmann (2018) proposed a multilevel factor scores analysis using expected a posteriori (EAP) estimates of factor scores. However, in the current state of development, his method relies on rather strict assumptions, such as all factor loadings need to be equal to each other and all error variances at the within and between level need to be equal.

The approach of Zitzmann (2018) was based on a method suggested by Croon and van Veldhoven (2007), where EAP-estimates were used instead of group means to aggregate scores measured at the within level to the between level. Using these EAP-estimates made

it possible to use variables measured at the within level to predict variables measured at the between level. While the method is also suggested by Croon, it differs from the method we are suggesting in several ways. First, Croon and van Veldhoven (2007) limited themselves to micro-macro settings, where variables from the within level are used to predict variables from the between level. This also means their illustrations were limited to linear regression settings with only one dependent variable. However, whether the Croon and van Veldhoven method itself can be extended to multivariate settings (i.e., two or more dependent variables) remains subject to future research. Our method is general and can be used for all settings, including the macro-micro setting, where variables measured at the between level are used to predict variables at the within level, multivariate regressions, and path analyses. Second, in the approach of Croon and van Veldhoven (2007), the variables were observed variables at the within level. Only the group averages at the between level were considered to be latent. Their method cannot deal with latent variables at the within level, while our method can. In this paper, we propose a corrected multilevel factor score analysis, that is more flexible than the method of Croon and van Veldhoven (2007) and its extension by Zitzmann (2018).

While stepwise estimation methods have many advantages, as discussed above, they are also often critized for potentially being less efficient than MLE, since they are limited information approaches (e.g. Wooldridge, 2002). However, for the single level setting, it was shown that the Croon method is just as efficient as MLE (Devlieger et al., 2016). Croon and van Veldhoven (2007) also expected that their approach would probably not be much less efficient than MLE. Lüdtke et al. (2008) tested this expectation in a simulation study and found almost identical root means squared errors (RMSE) for MLE and the method of Croon and van Veldhoven (2007). They only found a difference in one condition, namely the condition with small sample sizes at both the between (50 clusters) and within level (10 individuals per cluster), and a small intraclass correlation

coefficient (ICC). We would expect similar results for our method. The outline of the paper is as follows. First, we will give an overview of the within-between framework and of the method of Croon in the single-level setting. Next, we will propose a new statistical procedure that applies the method of Croon to the within-between framework. This will enable us to perform multilevel factor score regressions and multilevel factor score path analyses. Then, a simulation study will be presented to compare the performance of the extended Croon method to standard multilevel SEM using MLE. Finally, the method will be applied to data from a Flemish research project conducted in 2012, named MICTIVO.

5.2 The within-between framework

To extend the method of Croon to the multilevel setting, we will use the within-between framework. For notational simplicity, we will assume centered data and ignore the mean structure, but our method is not restricted to this setting. In the within-between framework for the two-level setting with subjects within clusters, the data are orthogonally decomposed into a within–group and a between–group component:

$$\mathbf{y}_T = \mathbf{y}_w + \mathbf{y}_b, \tag{5.1}$$

with $\mathbf{y}_b$ equal to the cluster mean and $\mathbf{y}_w$ equal to the individual deviation of a subject from the cluster mean. Both components are assumed to follow a normal distribution with mean zero and a covariance matrix:

$$\mathbf{y}_w \sim N(\mathbf{0}, \mathbf{\Sigma}_w) \tag{5.2}$$

$$\mathbf{y}_b \sim N(\mathbf{0}, \mathbf{\Sigma}_b), \tag{5.3}$$

where $\mathbf{\Sigma}_w$ represents the (co)variation within the clusters, while $\mathbf{\Sigma}_b$ represents the (co)variation between the clusters. In multilevel SEM, two structural equation models, one within model and one

between model, are jointly estimated for respectively $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_b$. The within model, containing a structural and a measurement part, can be written as

$$\begin{aligned} \boldsymbol{\eta}_w &= \mathbf{B}_w\boldsymbol{\eta}_w + \boldsymbol{\zeta}_w && (5.4) \\ \mathbf{y}_w &= \boldsymbol{\Lambda}_w\boldsymbol{\eta}_w + \boldsymbol{\epsilon}_w. && (5.5) \end{aligned}$$

The between model, also containing a structural and measurement part, can be written as

$$\begin{aligned} \boldsymbol{\eta}_b &= \mathbf{B}_b\boldsymbol{\eta}_b + \boldsymbol{\zeta}_b && (5.6) \\ \mathbf{y}_b &= \boldsymbol{\Lambda}_b\eta_b + \boldsymbol{\epsilon}_b, && (5.7) \end{aligned}$$

with subscripts 'w' and 'b' denoting the within and between level respectively, $\mathbf{B}_w$ and $\mathbf{B}_b$ are the matrices of regression coefficients, $\boldsymbol{\zeta}_w$ and $\boldsymbol{\zeta}_b$ the residual error terms, $\boldsymbol{\Lambda}_w$ and $\boldsymbol{\Lambda}_b$ contain the factor loadings, $\mathbf{y}_w$ a vector of indicators measuring $\boldsymbol{\eta}_w$, $\mathbf{y}_b$ a random cluster-specific intercept, and $\boldsymbol{\epsilon}_w$ and $\boldsymbol{\epsilon}_b$ the vectors of measurement error variables.
The formulas for the respective model-implied covariance matrices are:

$$\begin{aligned} \boldsymbol{\Sigma}_w &= \boldsymbol{\Lambda_w}(\boldsymbol{I} - \mathbf{B_w})^{-1}\boldsymbol{\Psi_w}(\boldsymbol{I} - \boldsymbol{B_w})^{-1'}\boldsymbol{\Lambda}_{\boldsymbol{w}}^{'} + \boldsymbol{\Theta_w}, && (5.8) \\ \boldsymbol{\Sigma}_b &= \boldsymbol{\Lambda_b}(\boldsymbol{I} - \boldsymbol{B_b})^{-1}\boldsymbol{\Psi_b}(\boldsymbol{I} - \boldsymbol{B_b})^{-1'}\boldsymbol{\Lambda}_{\boldsymbol{b}}^{'} + \boldsymbol{\Theta_b}, && (5.9) \end{aligned}$$

with $\boldsymbol{I}$ the identity matrix, $\boldsymbol{\Psi_w}$ and $\boldsymbol{\Psi_b}$ are matrices containing the variances of the latent variables, while $\boldsymbol{\Theta_w}$ and $\boldsymbol{\Theta_b}$ contain the residual variances of the indicators.

5.3 The Croon method in the single-level setting

The method of Croon uses the variances and covariances of factor scores to estimate the variances and covariances of the true latent variable scores. Consider measurement block 'k':

$$\mathbf{y}_k = \boldsymbol{\Lambda}_k \boldsymbol{\eta}_k + \boldsymbol{\epsilon}_k \tag{5.10}$$

where $\mathbf{y}_k$ are vectors of mean-centered observed indicators measuring $\boldsymbol{\eta}_k$, $\boldsymbol{\Lambda}_k$ are matrices containing the factor loadings and $\boldsymbol{\epsilon}_k$ are the vectors of measurement error variables. In many cases, a measurement block contains only a single latent variable. In that case, $\boldsymbol{\Lambda}_k$ is a one-column matrix. However, when latent variables are somehow linked (e.g equality constraints, correlated error terms,...), a measurement block can contain multiple latent variables.
After fitting the measurement block using confirmatory factor analyses (CFA), the factor scores can be calculated as follows:

$$\mathbf{F}_k \quad = \mathbf{A}_k \mathbf{y}_k \tag{5.11}$$

with $\mathbf{A}_k$ the factor score matrix. Several predictors can be used to compute the factor score matrices. Our method will work with every predictor, but in this paper the Regression predictor (Thomson, 1934; Thurstone, 1935) is used.
Croon established that one can estimate the variances and covariances of the true latent variable scores as follows:

$$\widehat{var(\eta_k)} = (\mathbf{A}_k\boldsymbol{\Lambda}_k)^{-1}[var(\mathbf{F}_k) - \mathbf{A}_k\boldsymbol{\Theta}_k\mathbf{A}_k^{'}]\ (\boldsymbol{\Lambda}_k^{'}\mathbf{A}_k^{'})^{-1} \tag{5.12}$$

$$\widehat{cov(\eta_k, \eta_m)} = (\mathbf{A}_k\boldsymbol{\Lambda}_k)^{-1} cov(\mathbf{F}_k, \mathbf{F}_m)\ (\boldsymbol{\Lambda}_m^{'}\mathbf{A}_m^{'})^{-1}, \tag{5.13}$$

where the indices 'k' and 'm' indicate two different measurement blocks with $\boldsymbol{\Theta}_k$ the covariance matrix of the measurement errors. For more details on the derivation of these correction formulas, the interested reader is referred to Croon (2002) and Devlieger et al. (2016). These formulas can be used to perform factor score regression or factor score path analysis. Devlieger and Rosseel (2017) suggested the following procedure:

1. Perform a separate factor analysis for each measurement block in the model and calculate the respective factor scores.

2. Calculate the variance–covariance matrix of the factor scores ($var(\mathbf{F}_k)$).

3. Estimate the true variances and covariances for all elements in this matrix ($\hat{\boldsymbol{\Sigma}}_\eta$) using the formulas of Croon.

4. Perform the analysis of interest, either linear regression (a) or path analysis (b).

   (a) In linear regression, regression parameters ($\boldsymbol{\gamma}$) can be defined as the covariance between the dependent and the independent variables, multiplied by the inverse of the true variance of the independent variable(s). Thus, the regression parameters can be calculated using the elements of interest from the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_\eta$. When $\boldsymbol{\eta}_m$ is a vector containing the independent variables and $\boldsymbol{\eta}_k$ a vector with the dependent variables, the regression parameters can be estimated as follows:

   $$\boldsymbol{\gamma} = \widehat{cov(\boldsymbol{\eta}_k, \boldsymbol{\eta}_m)} \widehat{var(\boldsymbol{\eta}_m)}^{-1}. \tag{5.14}$$

   (b) Path analysis is based on the covariance matrix of all variables. Thus, $\hat{\boldsymbol{\Sigma}}_\eta$ can be used as the input covariance matrix.

5. Standard errors of the parameters of interest can be obtained by an analytical approach where we take the stepwise nature of the procedure into account (Bakk, Oberski, & Vermunt, 2014).

5.4 The Croon method in the multilevel setting

The same principle can be used to perform a multilevel factor score analysis. This analysis can be a multilevel regression or a multilevel path analysis, depending on the structural part of the multilevel SEM. When using factor scores to represent the latent variables,

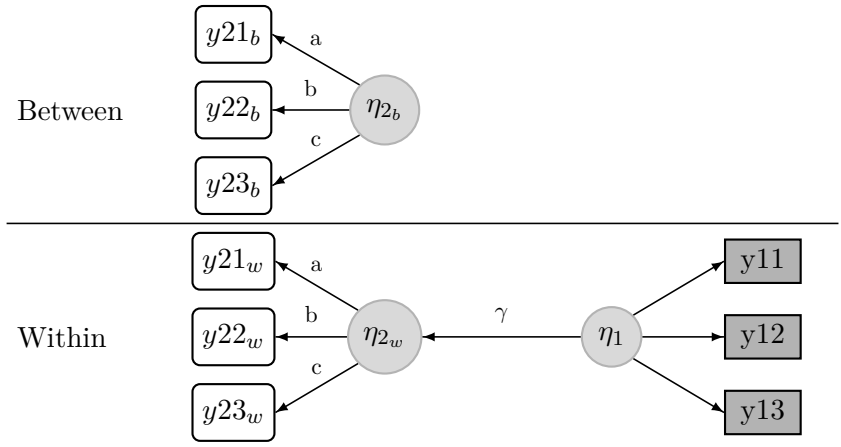


Figure 5.1 Multilevel SEM model that reduces to a multilevel regression with a random intercept, when using factor scores.

the model can either reduce to the multilevel regression setting or reduce to a multilevel path analysis. Figure 5.1 is an example of a multilevel SEM model that reduces to a multilevel regression with a fixed slope and a random intercept. Figure 5.2 is an example of a multilevel SEM model that reduces to a multilevel path analysis with three regressions on the between level and three regressions on the within level. Our procedure is based on the procedure for single level FSR or path analysis. However, some steps are adjusted to the multilevel setting.

1. The first step is to perform a multilevel CFA for each measurement block separately and calculate the within factor scores. The measurement models are

$$\mathbf{y}_w = \boldsymbol{\Lambda}_w \boldsymbol{\eta}_w + \boldsymbol{\epsilon}_w \tag{5.15}$$

$$\mathbf{y}_b = \boldsymbol{\Lambda}_b \boldsymbol{\eta}_b + \boldsymbol{\epsilon}_b. \tag{5.16}$$

Based on this factor model, the within factor scores $\mathbf{F}_w$ and

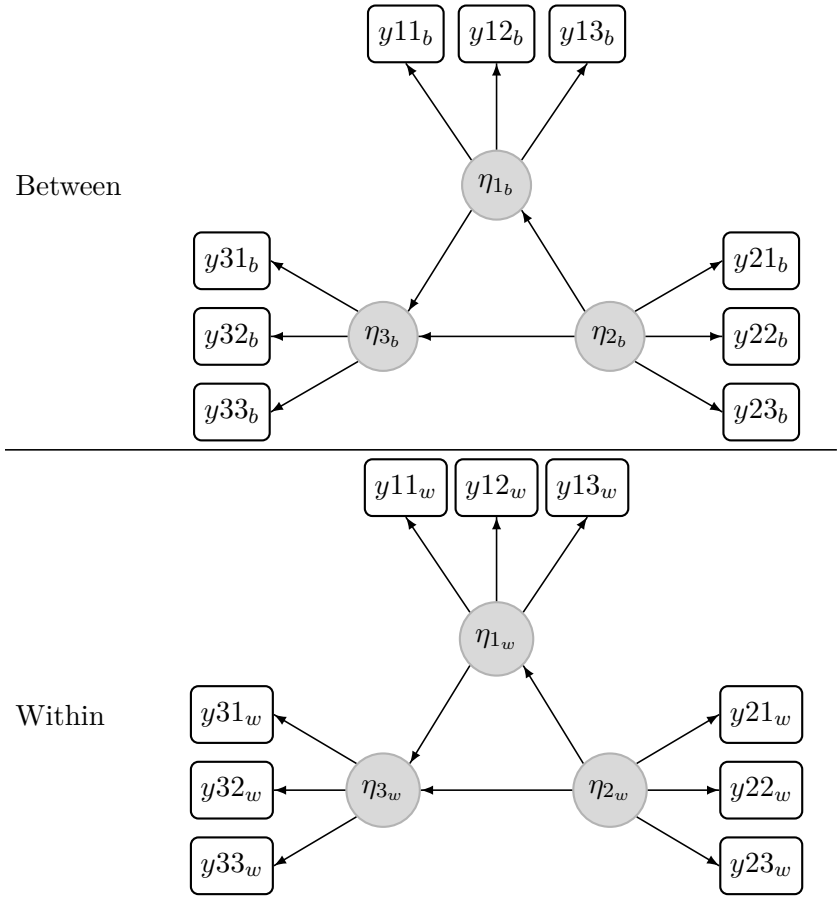


Figure 5.2 Multilevel SEM model that reduces to a multilevel path analysis when using factor scores.

the between factor scores $\mathbf{F}_b$ can be calculated as follows

$$\mathbf{F}_w = \mathbf{A}_w \mathbf{y}_w \tag{5.17}$$
$$\mathbf{F}_b = \mathbf{A}_b \mathbf{y}_b. \tag{5.18}$$

with $\mathbf{A}_w$ and $\mathbf{A}_b$ the within and between factor score matrix, $\mathbf{y}_w$ the observed mean-centered within data and $\mathbf{y}_b$ the observed cluster means. The exact computation of $\mathbf{A}_w$ and $\mathbf{A}_b$ depends on the predictor that is used. Note that the observed cluster means are usually considered unreliable (Lüdtke et al., 2008). However, the bias that would normally be caused by using aggregated means is taken into account by the bias formulas in step 3.

2. In the second step, two covariance matrices of the factor scores need to be calculated, namely the between cluster covariance matrix $\boldsymbol{S}_{Fb}$ and the pooled within cluster covariance matrix $\boldsymbol{S}_{Fw}$,

$$\boldsymbol{S}_{Fb} = \frac{1}{J-1} \Sigma_{j=1}^{J} (F_{b_j} - \bar{F}_b)(F_{b_j} - \bar{F}_b)' \tag{5.19}$$
$$\boldsymbol{S}_{Fw} = \frac{1}{N-J} \Sigma_{j=1}^{J} \Sigma_{i=1}^{n_j} (F_{w_{ij}} - F_{b_j})(F_{w_{ij}} - F_{b_j})' \tag{5.20}$$

with subscripts i and j denoting the subjects and clusters, N the total sample size, J the number of clusters and n_j the cluster size of cluster j.

3. The elements in both $\boldsymbol{S}_{Fb}$ and $\boldsymbol{S}_{Fw}$ are biased. Thus, in this third step, all elements need to be corrected using the formulas of Croon. For the within level, the formulas are the same as for the single-level setting:

$$var(\boldsymbol{\eta}_w) = (\mathbf{A}_w \boldsymbol{\Lambda}_w)^{-1} [var(\mathbf{F}_w) - \mathbf{A}_w \boldsymbol{\Theta}_w \mathbf{A}_w'] (\boldsymbol{\Lambda}_w' \mathbf{A}_w')^{-1} \tag{5.21}$$
$$cov(\boldsymbol{\eta}_{wk}, \boldsymbol{\eta}_{wm}) = (\mathbf{A}_{wk} \boldsymbol{\Lambda}_{wk})^{-1} cov(\mathbf{F}_{wk}, \mathbf{F}_{wm}) (\boldsymbol{\Lambda}_{wm}' \mathbf{A}_{wm}')^{-1}, \tag{5.22}$$

However, for the between level, the formulas need to be adjusted. In Appendix A and B, we derive the following formulas:

$$var(\boldsymbol{\eta}_b) = (\mathbf{A}_b\boldsymbol{\Lambda}_b)^{-1}[var(\boldsymbol{F}_b) - \mathbf{A}_b\boldsymbol{\Theta}_b\mathbf{A}_b^{'} - \frac{1}{n_j}\mathbf{A}_b var(\mathbf{y}_w^*)\mathbf{A}_b^{'}](\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'})^{-1}, \tag{5.23}$$

and

$$cov(\boldsymbol{\eta}_{bk}, \boldsymbol{\eta}_{bm}) = (\mathbf{A}_{bk}\boldsymbol{\Lambda}_{bk})^{-1}[cov(F_{bk}, F_{bm}) - \frac{1}{n_j}\mathbf{A}_{bk} cov(\mathbf{y}_{wk}^*, \mathbf{y}_{wm}^*)\mathbf{A}_{bm}^{'}](\boldsymbol{\Lambda}_{bm}^{'}\mathbf{A}_{bm}^{'})^{-1}, \tag{5.24}$$

This results in two new estimated covariance matrices, respectively $\hat{\boldsymbol{\Sigma}}_{\eta_w}$ and $\hat{\boldsymbol{\Sigma}}_{\eta_b}$.

4. In the fourth step, the two new estimated covariance matrices can be used as input for two separate structural equation models, one for the within level and one for the between level. If the structural models of the within and between levels are linked (e.g. equality constraints across levels), the covariance matrices can be used as input for a multigroup SEM.

5. Standard errors of the parameters of interest can be obtained by an analytical approach that takes the stepwise nature of the procedure into account (Bakk et al., 2014).

There are three aspects concerning this procedure that are important to highlight. Firstly, step 4 can only be performed if the software accepts sample statistics as input. This is true for most SEM packages. However, regression based software (such as MLwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2005) or the R-package 'lme4' (Bates, Maechler, Bolker, & Walker, 2015)) only accepts raw data. In appendix C, it is described how these data can be obtained. Secondly, as is shown in appendices A en B, the

formulas can be rewritten to avoid calculating the factor scores and their covariance matrices explicitly. Note that this is also true for the within formulas. The covariance matrices of the latent variables can be estimated using only the covariance matrices of the indicator items, the factor score matrices, and the factor loadings and error terms of the items. Thirdly, all formulas above are formulated at the population level. This implies that the method will certainly work with large samples. To evaluate the finite sample performance of the new estimation method, a simulation study was used.

5.5 Simulation study

The first goal of this study was to see if the new method results in unbiased estimates of the regression parameters and variance terms. The second goal was to compare the method to standard multilevel SEM. We studied which method performs better when there are a small number of clusters or when there are misspecifications in the model with regard to bias, MSE, and coverage.

5.5.1 Data generation

The model in Figure 5.3 was used to simulate the data in R (R Core Team, 2016) using the MASS-package (Venables & Ripley, 2002). The between–cluster data were simulated first, by drawing from a multivariate normal distribution with all means 0 and a covariance matrix Σ_b. Σ_b was calculated using formula (5.9). All four latent variables were measured using 5 indicators with factor loadings 1, 0.8, 0.9, 0.85, and 0.75. The regression parameters were set to $\gamma_1 = 0.7$, $\gamma_2 = 0.5$,$\gamma_3 = 0.4$, $\gamma_4 = 0.5$, and $\gamma_5 = 0.8$. The between variances of $\boldsymbol{\eta}_1$ and the residual between variance of $\boldsymbol{\eta}_2$, $\boldsymbol{\eta}_3$, and $\boldsymbol{\eta}_4$ were all set to 0.9. The residual between variances of the indicator items were set to $\Theta_{b_{ki}} = \frac{var(\eta_{b_k})(1-CD_{ki})}{CD_{ki}}$ with CD_{ki} the coefficients of determination. The index 'i' is used to refer to the items, while index 'k' refers to the latent variables. All CD_{ki} were set to be equal and were varied between 0.5,0.7, and 0.9. The

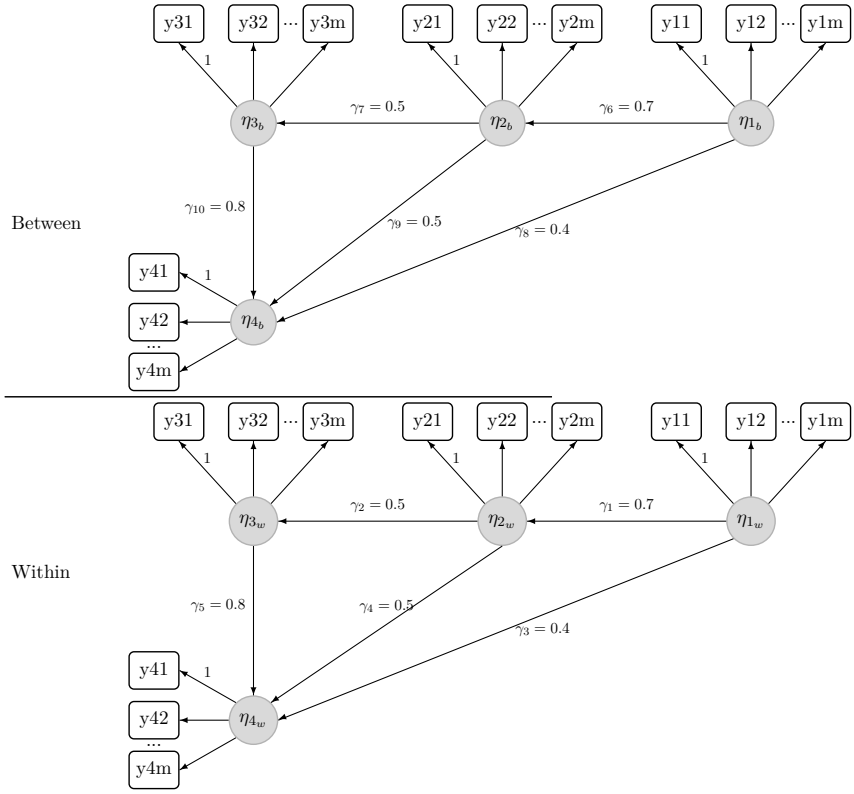


Figure 5.3 The ground truth model for the simulation study.

number of clusters was set to 50 or 100. These can be considered as small sample sizes for the between level. We chose these to be able to evaluate if the method of Croon can handle these low number of clusters better than MLE.
Then, the within data were simulated per cluster, by drawing from a multivariate normal distribution using the between data as the mean of the cluster and a covariance matrix Σ_w. Σ_w was calculated using formula (5.8). The within factor loadings were set to be equal to the between factor loadings. The regression parameters were set to $\gamma_6 = 0.7$, $\gamma_7 = 0.5$, $\gamma_8 = 0.4$, $\gamma_9 = 0.5$, and $\gamma_{10} = 0.8$. The within variances of $\boldsymbol{\eta}_1$ and the residual within variance of $\boldsymbol{\eta}_2$, $\boldsymbol{\eta}_3$, and $\boldsymbol{\eta}_4$ were all set to $\frac{var(\eta_{b_k})\times(1-ICC)}{ICC}$ with ICC the intraclass correlation coefficient, which indicates the amount of variability on the between level. The ICC was varied between 0.10 and 0.25. The within residual variances were set to $\Theta_{w_{ki}} = \frac{var(\eta_{w_k})(1-CD_{ki})}{CD_{ki}}$. Since all CD_{ki} were set to be equal, they will be referred to as CD. The cluster sizes were randomly drawn with replacement from a uniform distribution between 5 and 15, 25 and 35, or 100 and 110, resulting in unbalanced data. All together, this created 36 experimental conditions.

5.5.2 Analysis

For each experimental condition, 1000 datasets were generated. All datasets were analyzed using three methods, namely uncorrected multilevel factor score path analysis, multilevel factor score path analysis with Croon corrections and Multilevel SEM using MLE. Multilevel SEM using MLE was performed using Mplus, version 8 (L. Muthén & Muthén, 2017), which uses the EM algorithm by default. For the two multilevel factor scores path analyses, Mplus (L. Muthén & Muthén, 2017) was used to perform the CFAs, calculate the factor scores and perform the multilevel path analyses. The Croon corrections were done using our own written routines and the MASS-package (Venables & Ripley, 2002) was used to generate the corrected datasets.

For each method, three models were analysed, namely a correctly specified model, a model containing a misspecification in the structural part of the model, and a model containing a misspecification in the measurement part of the model. For the misspecification in the structural part of the model, regression parameters γ_3 and γ_8 were left out of the model. This should only have an effect on γ_4 and γ_9. The other parameters should remain unaffected. For the misspecification of the measurement model, indicator y21 was loaded on variable η_3 instead of η_2, both at the within and between level.
For each experimental condition, four performance criteria per method were calculated, namely the proportion of successful replications, the relative bias, MSE, and coverage rate. The latter three criteria were calculated for each regression coefficient and variance term, both at the within and the between level.

5.5.3 Results

Proportion of successful replications

Figure 5.4 shows us that the number of successful replications was the highest for the uncorrected factor score path analysis (FSPA). There were mainly unsuccessful replications in the weaker conditions, such as small number of clusters, small cluster sizes, or a small ICC. In Table 5.1, an overview of the error and warning messages from Mplus is given for the three methods in the weakest condition (nb=50, cluster size = 5-15, CD=0.5 and ICC = 0.10) in the correctly specified model. It can be seen that the problems were mainly due to one of the factor models having a residual covariance matrix or a latent variable covariance matrix that was not positive definite.
Of course, the method of Croon suffered from exactly the same problems, but also had another problem: additional replications did not converge due to covariance matrices at the between level that were not positive definite after the correction was applied.
When comparing the Croon method with MLE, it is striking that

the Croon method had higher or at least equal proportions of successful replications in all of the 36 conditions for all three models. This includes the correctly specified model, indicating that the non-convergence of MLE was not always an indication of misspecification. For MLE, unsuccessful replications were mainly caused by the latent variable covariance matrix not being positive definite or the first-order derivative product matrix being non-positive definite. Some additional replications were unsuccessful due to the residual covariance matrix not being positive definite or an ill-conditioned fisher information matrix. Note that none of the unsuccessful replications were due to the model not converging because the number of iterations was exceeded. Also note that the Croon method almost always converged when the number of clusters reached 100 or when the clusters consisted of at least 25 elements, while MLE still had convergence problems in these settings.

Bias

Only the results of the correct model and the model with structural misspecifications are discussed, because no interesting differences were found between the Croon method and MLE in the model with a misspecification in the measurement model. This will also be the case in the following sections.

Regression parameters: Figure 5.5a shows the results for the bias of the regression parameters on the within level, while Figure 5.5b shows the bias on the between level. For the model with a structural misspecification, regression paramaters γ_3 and γ_4 were left out of the figure, since the first was not included in the model and the latter is expected to be biased for all methods. The full results for this model can be found in Tables 5.2 and 5.3.

As expected, the uncorrected multilevel FSPA resulted in biased regression parameters on the within and between level, for all conditions and models (Figure 5.5). The Croon method showed no bias when the correctly specified model is used, both on the within and between level. It did show a slight bias on the between level when

Table 5.1 An overview of the error and warning messages, for the correctly specified model with 50 clusters, a cluster size between 5 and 15, a coefficient of determination of 0.5, and a intra-class correlation of 0.10. An overview is given for the three methods, namely factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE)

	FSPA	Croon	MLE
Warning: the residual covariance matrix (theta) is not positive definite.	157	157	80
Warning: the latent variable covariance matrix (psi) is not positive definite.	46	46	287
The corrected between covariance matrix is not positive definite.	/	324	/
Warning: the model estimation had reached a saddle point or a point where the observed and the expected information matrices do not match.	17	1	0
The standard errors of the model parameter estimates may not be trustworthy for some parameters due to a non-positive definite first-order derivative product matrix.	3	0	260
The model estimation did not terminate normally due to an ill-conditioned fisher information matrix.	1	1	47
The model estimation did not terminate normally due to an error in the computation.			5
The loglikelihood decreased in the last EM iteration.			4
Total	224	529	683

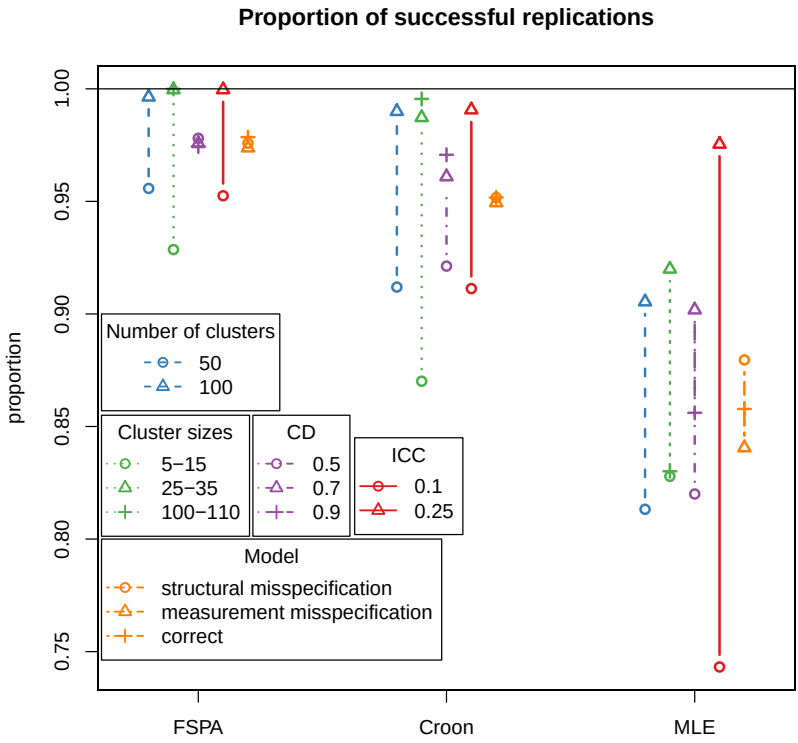


Figure 5.4 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the proportion of successful replications for the three methods; factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE).

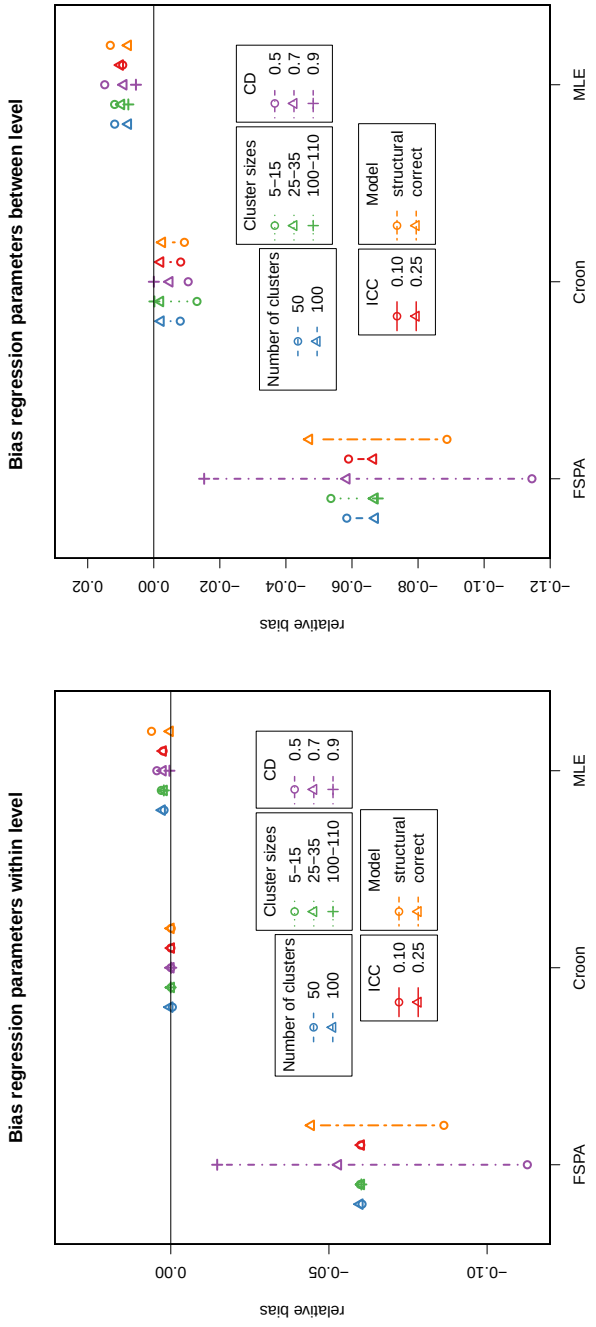
there is a misspecification in the structural model. However, this bias was only found in the weakest condition with only 50 clusters, a cluster size between 5 and 15, CD of 0.5, and ICC of 0.10. MLE, on the other hand was only unbiased on the within level, when a correctly specified model was used. On the between level, MLE was biased, even when the model was correctly specified. A misspecification in the model lead to bias on the within and between level.

Variance terms: Figure 5.6 shows the bias in the variance terms of the endogeneous variables. The residual variance of η_4 was expected to be biased, since the regression of η_2 on η_4 was dropped from the model. Therefore the results of this parameters were again left out of the figure. The full results can be found in Tables 5.4 and 5.5. The estimated residual within and between variances of the endogeneous latent variables (η_2, η_3, and η_4) were highly biased for the uncorrected FSPA. The within variance of the Croon method was unbiased in all settings. On the between level, there was a small bias that disappeared when the coefficient of determination increased. For MLE, the within variance was almost unbiased when the model was correctly specified, but was biased when the model was misspecified. MLE also had a bias in the between variance terms that was larger than the bias of the Croon method.

MSE

Overall, the Croon method clearly outperformed MLE with regard to bias in the regression parameters and variance terms. However, it is imporant to also take the efficiency of methods into account. For this reason, we will also look at the MSE of all methods.

Regression parameters: The results with regard to MSE were very similar to the results found with regard to the bias (Figure 5.7). The high bias of the uncorrected multilevel FSPA also resulted in a high MSE. When a correctly specified model was used, the MSE of the Croon method, and MLE was very similar on the within level, but higher for MLE on the between level. When the model contained a



(a) The within regression parameters.

(b) The between regression parameters.

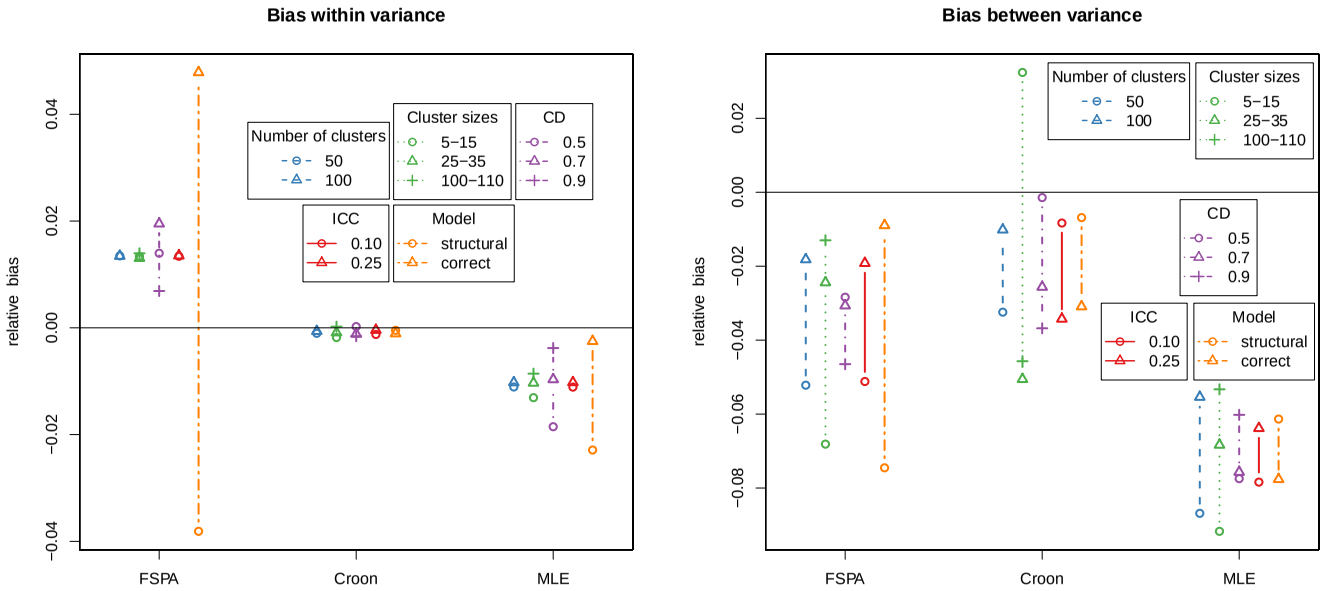
Figure 5.5 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the bias of the regression parameters for factor score path analysis (FSPA), the Croon method, and maximum likelihood estimation (MLE).

Table 5.2 The estimated within regression parameters for the model with a misspecification in the structural part. Results are given for different values of intra-class correlations (ICC), cluster sizes (nw), number of clusters (nb), coefficient of determination (CD), and the three methods, namely factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE).

				$\gamma_1 = 0.7$			$\gamma_2 = 0.5$			$\gamma_4 = 0.5$			$\gamma_5 = 0.8$		
ICC	nw	nb	CD	FSPA	Croon	MLE	FSPA	Croon	MLE	FSPA	Croon	MLE	FSPA	Croon	MLE
0.10	5–	50	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.67	0.69	0.75	0.75	0.80	0.77
	15		0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.78
			0.9	0.68	0.70	0.71	0.48	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
		100	0.5	0.56	0.70	0.74	0.40	0.50	0.51	0.67	0.69	0.75	0.75	0.80	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
	25–	50	0.5	0.55	0.70	0.74	0.39	0.50	0.50	0.67	0.69	0.75	0.75	0.80	0.77
	35		0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
		100	0.5	0.55	0.70	0.74	0.39	0.50	0.51	0.66	0.69	0.75	0.75	0.80	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
	100–	50	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.66	0.69	0.75	0.75	0.80	0.77
	110		0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
		100	0.5	0.55	0.70	0.74	0.39	0.50	0.50	0.66	0.69	0.75	0.75	0.80	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.79	0.80	0.80
0.25	5–	50	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.67	0.69	0.75	0.75	0.80	0.77
	15		0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
		100	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.67	0.69	0.75	0.75	0.80	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
	25–	50	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.66	0.69	0.75	0.75	0.80	0.77
	35		0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
		100	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.66	0.69	0.75	0.75	0.80	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
	100–	50	0.5	0.55	0.70	0.74	0.40	0.50	0.51	0.66	0.69	0.75	0.75	0.80	0.77
	110		0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.80	0.80	0.80
		100	0.5	0.55	0.70	0.74	0.39	0.50	0.50	0.66	0.69	0.75	0.75	0.80	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.69	0.72	0.78	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.69	0.69	0.70	0.79	0.80	0.80

Table 5.3 The estimated between regression parameters for the model with a misspecification in the structural part. Results are given for different values of intra-class correlations (ICC), cluster sizes (nw), number of clusters (nb), coefficient of determination (CD), and the three methods, namely factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE).

				$\gamma_1 = 0.7$			$\gamma_2 = 0.5$			$\gamma_4 = 0.5$			$\gamma_5 = 0.8$		
ICC	nw	nb	CD	FSPA	Croon	MLE	FSPA	Croon	MLE	FSPA	Croon	MLE	FSPA	Croon	MLE
0.10	5–	50	0.5	0.57	0.67	0.74	0.50	0.49	0.51	0.82	0.68	0.69	0.76	0.79	0.80
	15		0.7	0.59	0.69	0.71	0.48	0.49	0.52	0.77	0.68	0.67	0.78	0.81	0.83
			0.9	0.63	0.70	0.69	0.51	0.50	0.52	0.64	0.68	0.71	0.78	0.80	0.79
		100	0.5	0.53	0.70	0.72	0.42	0.50	0.51	0.74	0.70	0.76	0.69	0.78	0.76
			0.7	0.58	0.69	0.71	0.45	0.49	0.50	0.70	0.69	0.73	0.76	0.80	0.80
			0.9	0.63	0.71	0.68	0.48	0.50	0.50	0.70	0.69	0.69	0.79	0.80	0.81
	25–	50	0.5	0.57	0.69	0.75	0.41	0.49	0.52	0.66	0.66	0.73	0.75	0.80	0.80
	35		0.7	0.63	0.69	0.72	0.45	0.50	0.51	0.67	0.69	0.74	0.76	0.80	0.79
			0.9	0.68	0.70	0.71	0.48	0.50	0.50	0.67	0.68	0.69	0.80	0.81	0.80
		100	0.5	0.56	0.70	0.75	0.40	0.49	0.51	0.65	0.69	0.76	0.72	0.79	0.77
			0.7	0.62	0.69	0.72	0.45	0.50	0.51	0.68	0.70	0.74	0.76	0.80	0.78
			0.9	0.69	0.71	0.72	0.49	0.50	0.50	0.68	0.69	0.69	0.79	0.80	0.80
	100–	50	0.5	0.56	0.69	0.74	0.40	0.50	0.51	0.63	0.68	0.74	0.74	0.81	0.80
	110		0.7	0.63	0.70	0.72	0.45	0.50	0.51	0.67	0.69	0.74	0.76	0.80	0.78
			0.9	0.68	0.70	0.72	0.49	0.50	0.51	0.68	0.69	0.69	0.79	0.80	0.80
		100	0.5	0.56	0.70	0.74	0.40	0.50	0.51	0.65	0.70	0.75	0.72	0.79	0.77
			0.7	0.63	0.70	0.72	0.44	0.49	0.50	0.67	0.69	0.72	0.77	0.80	0.80
			0.9	0.69	0.71	0.73	0.48	0.50	0.50	0.69	0.69	0.70	0.79	0.80	0.80
0.25	5–	50	0.5	0.57	0.67	0.75	0.42	0.49	0.53	0.65	0.68	0.76	0.73	0.79	0.78
	15		0.7	0.63	0.69	0.73	0.45	0.49	0.51	0.67	0.68	0.73	0.77	0.81	0.80
			0.9	0.69	0.70	0.71	0.49	0.50	0.51	0.67	0.68	0.69	0.79	0.80	0.80
		100	0.5	0.56	0.70	0.75	0.40	0.50	0.52	0.65	0.70	0.77	0.71	0.78	0.77
			0.7	0.63	0.69	0.72	0.44	0.49	0.50	0.68	0.69	0.73	0.77	0.80	0.80
			0.9	0.69	0.71	0.72	0.48	0.50	0.50	0.69	0.69	0.70	0.79	0.80	0.80
	25–	50	0.5	0.56	0.69	0.75	0.40	0.50	0.52	0.63	0.66	0.74	0.73	0.82	0.80
	35		0.7	0.63	0.70	0.73	0.45	0.50	0.51	0.67	0.69	0.73	0.76	0.79	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.68	0.68	0.69	0.79	0.80	0.80
		100	0.5	0.56	0.70	0.75	0.40	0.50	0.51	0.64	0.69	0.76	0.72	0.79	0.77
			0.7	0.63	0.70	0.72	0.45	0.50	0.50	0.68	0.70	0.73	0.76	0.80	0.79
			0.9	0.69	0.71	0.72	0.48	0.50	0.50	0.68	0.69	0.70	0.79	0.80	0.80
	100–	50	0.5	0.56	0.69	0.74	0.40	0.50	0.51	0.63	0.68	0.74	0.73	0.81	0.79
	110		0.7	0.63	0.70	0.73	0.45	0.50	0.51	0.67	0.69	0.73	0.76	0.80	0.79
			0.9	0.68	0.70	0.71	0.49	0.50	0.50	0.68	0.69	0.70	0.79	0.80	0.80
		100	0.5	0.56	0.70	0.74	0.40	0.50	0.51	0.65	0.70	0.76	0.72	0.79	0.77
			0.7	0.63	0.70	0.72	0.45	0.49	0.50	0.67	0.69	0.72	0.77	0.80	0.79
			0.9	0.69	0.71	0.71	0.48	0.50	0.50	0.69	0.69	0.70	0.79	0.80	0.79



(a) Results for the within variance terms.

(b) Results for the between variance terms.

Figure 5.6 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the bias of the variance terms for the three methods; factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE).

Table 5.4 The estimated residual within variances of the endogeneous variables for the model with a misspecification in the structural part. Results are given for different values of intra-class correlations (ICC), cluster sizes (nw), number of clusters (nb), coefficient of determination (CD), and the three methods, namely factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE). The true values are 8.1 (ICC=0.10) and 2.7 (ICC=0.25).

				η_4			η_3			η_2		
ICC	nw	nb	CD	FSPA	Croon	MLE	FSPA	Croon	MLE	FSPA	Croon	MLE
0.10	5–	50	0.5	12.12	8.89	8.11	7.31	8.06	8.10	7.62	8.15	7.45
	15		0.7	10.63	8.87	8.51	7.81	8.05	8.05	8.03	8.12	7.77
			0.9	9.44	8.93	8.81	8.02	8.07	8.07	8.07	8.05	7.96
		100	0.5	12.24	8.99	8.26	7.30	8.11	8.08	7.59	8.08	7.38
			0.7	10.72	9.00	8.63	7.77	8.07	8.06	7.94	8.06	7.74
			0.9	9.45	8.95	8.85	8.03	8.10	8.09	8.10	8.09	8.03
	25–	50	0.5	12.22	8.98	8.30	7.29	8.08	8.05	7.58	8.10	7.43
	35		0.7	10.70	8.98	8.64	7.77	8.07	8.07	7.95	8.09	7.77
			0.9	9.47	8.96	8.86	8.03	8.10	8.09	8.08	8.09	8.01
		100	0.5	12.19	8.96	8.29	7.29	8.10	8.09	7.59	8.12	7.44
			0.7	10.69	8.95	8.62	7.78	8.08	8.08	7.97	8.10	7.79
			0.9	9.47	8.96	8.87	8.03	8.09	8.09	8.07	8.09	8.00
	100–	50	0.5	12.20	8.97	8.32	7.29	8.11	8.10	7.58	8.11	7.46
	110		0.7	10.70	8.96	8.64	7.80	8.10	8.10	7.96	8.10	7.78
			0.9	9.49	8.97	8.88	8.03	8.10	8.11	8.08	8.09	8.00
		100	0.5	12.21	8.98	8.33	7.29	8.10	8.09	7.59	8.11	7.46
			0.7	10.70	8.97	8.64	7.81	8.11	8.11	7.97	8.10	7.79
			0.9	9.48	8.97	8.88	8.03	8.10	8.10	8.09	8.10	8.01
0.25	5–	50	0.5	4.05	2.97	2.72	2.45	2.72	2.70	2.54	2.72	2.48
	15		0.7	3.55	2.97	2.86	2.60	2.70	2.70	2.66	2.71	2.60
			0.9	3.15	2.98	2.94	2.67	2.70	2.70	2.68	2.69	2.66
		100	0.5	4.08	3.01	2.77	2.43	2.70	2.69	2.52	2.70	2.47
			0.7	3.57	3.00	2.88	2.59	2.69	2.69	2.65	2.70	2.59
			0.9	3.15	2.98	2.95	2.67	2.69	2.69	2.69	2.70	2.67
	25–	50	0.5	4.07	3.00	2.77	2.44	2.71	2.70	2.53	2.70	2.48
	35		0.7	3.57	2.99	2.88	2.60	2.70	2.70	2.65	2.69	2.59
			0.9	3.16	2.99	2.96	2.67	2.69	2.69	2.69	2.70	2.67
		100	0.5	4.07	2.99	2.77	2.43	2.70	2.70	2.53	2.70	2.48
			0.7	3.56	2.98	2.88	2.60	2.70	2.70	2.66	2.70	2.60
			0.9	3.16	2.99	2.96	2.67	2.70	2.70	2.69	2.70	2.67
	100–	50	0.5	4.06	2.99	2.77	2.43	2.70	2.70	2.52	2.70	2.48
	110		0.7	3.56	2.99	2.88	2.60	2.70	2.70	2.65	2.70	2.60
			0.9	3.16	2.99	2.96	2.68	2.70	2.70	2.69	2.70	2.67
		100	0.5	4.07	2.99	2.78	2.43	2.70	2.70	2.53	2.70	2.48
			0.7	3.57	2.99	2.88	2.60	2.70	2.70	2.66	2.70	2.60
			0.9	3.16	2.99	2.96	2.68	2.70	2.70	2.70	2.70	2.67

Table 5.5 The estimated residual between variances of the endogeneous variables for the model with a misspecification in the structural part. Results are given for different values of intra-class correlations (ICC), cluster sizes (nw), number of clusters (nb), coefficient of determination (CD), and the three methods, namely factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE). The true value is 0.9.

				η_4			η_3			η_2		
ICC	nw	nb	CD	FSPA	Croon	MLE	FSPA	Croon	MLE	FSPA	Croon	MLE
0.10	5–	50	0.5	0.99	1.26	0.79	0.75	1.09	0.86	0.76	1.16	0.94
	15		0.7	0.84	1.09	0.72	0.79	0.99	0.81	0.75	1.01	0.84
			0.9	0.74	0.96	0.70	0.79	0.92	0.78	0.76	0.93	0.78
		100	0.5	1.15	1.15	0.79	0.77	1.00	0.84	0.78	1.04	0.84
			0.7	0.99	1.03	0.77	0.81	0.96	0.83	0.81	0.98	0.81
			0.9	0.90	1.02	0.83	0.83	0.94	0.83	0.79	0.94	0.79
	25–	50	0.5	1.21	0.93	0.76	0.78	0.87	0.84	0.81	0.87	0.78
	35		0.7	1.11	0.92	0.83	0.83	0.85	0.84	0.83	0.84	0.80
			0.9	0.95	0.89	0.87	0.85	0.85	0.86	0.85	0.85	0.84
		100	0.5	1.32	0.98	0.84	0.80	0.88	0.87	0.83	0.88	0.80
			0.7	1.15	0.95	0.88	0.84	0.87	0.87	0.87	0.88	0.84
			0.9	1.02	0.95	0.94	0.87	0.87	0.88	0.88	0.87	0.87
	100–	50	0.5	1.29	0.94	0.83	0.80	0.87	0.89	0.83	0.87	0.78
	110		0.7	1.14	0.93	0.87	0.84	0.85	0.86	0.85	0.85	0.81
			0.9	0.99	0.91	0.91	0.86	0.85	0.87	0.87	0.85	0.88
		100	0.5	1.33	0.97	0.89	0.80	0.87	0.89	0.84	0.88	0.82
			0.7	1.17	0.96	0.91	0.85	0.88	0.88	0.88	0.89	0.85
			0.9	1.03	0.96	0.96	0.88	0.87	0.89	0.88	0.88	0.89
0.25	5–	50	0.5	1.23	1.00	0.77	0.80	0.92	0.85	0.83	0.94	0.81
	15		0.7	1.09	0.96	0.81	0.83	0.89	0.84	0.85	0.89	0.81
			0.9	0.96	0.94	0.87	0.85	0.88	0.85	0.85	0.87	0.84
		100	0.5	1.32	1.02	0.83	0.79	0.91	0.86	0.83	0.92	0.80
			0.7	1.15	0.99	0.88	0.84	0.90	0.86	0.87	0.92	0.84
			0.9	1.02	0.99	0.94	0.86	0.89	0.87	0.87	0.90	0.86
	25–	50	0.5	1.28	0.93	0.78	0.80	0.86	0.85	0.83	0.87	0.79
	35		0.7	1.14	0.94	0.87	0.84	0.85	0.86	0.84	0.84	0.81
			0.9	0.98	0.91	0.90	0.86	0.85	0.86	0.87	0.85	0.85
		100	0.5	1.33	0.98	0.87	0.80	0.88	0.87	0.84	0.88	0.81
			0.7	1.17	0.96	0.91	0.85	0.87	0.88	0.87	0.88	0.85
			0.9	1.03	0.96	0.95	0.87	0.87	0.88	0.88	0.88	0.87
	100–	50	0.5	1.30	0.93	0.81	0.80	0.87	0.87	0.83	0.87	0.80
	110		0.7	1.15	0.94	0.88	0.84	0.85	0.86	0.85	0.85	0.82
			0.9	0.99	0.92	0.91	0.86	0.85	0.86	0.87	0.86	0.86
		100	0.5	1.34	0.97	0.87	0.80	0.87	0.87	0.84	0.88	0.82
			0.7	1.17	0.96	0.92	0.85	0.88	0.88	0.88	0.89	0.86
			0.9	1.04	0.97	0.96	0.88	0.88	0.88	0.88	0.88	0.87

structural misspecification, the MSE was always slightly higher for MLE.

Variance terms: The MSE of the within variance term was very similar for MLE and the Croon method with the correct model, but lower for the Croon method for the structural misspecification (see Figure 5.8). On the between level, the MSE of the Croon method was slightly higher for the correct model, but slightly lower for the structural misspecification. The higher MSE in the correct model was caused by one condition, namely the weakest.

In conclusion, the Croon methods performed just as well as MLE or even outperformed MLE with regard to the MSE with an exception for the MSE of the between variance term in the correctly specified model in the weakest condition.

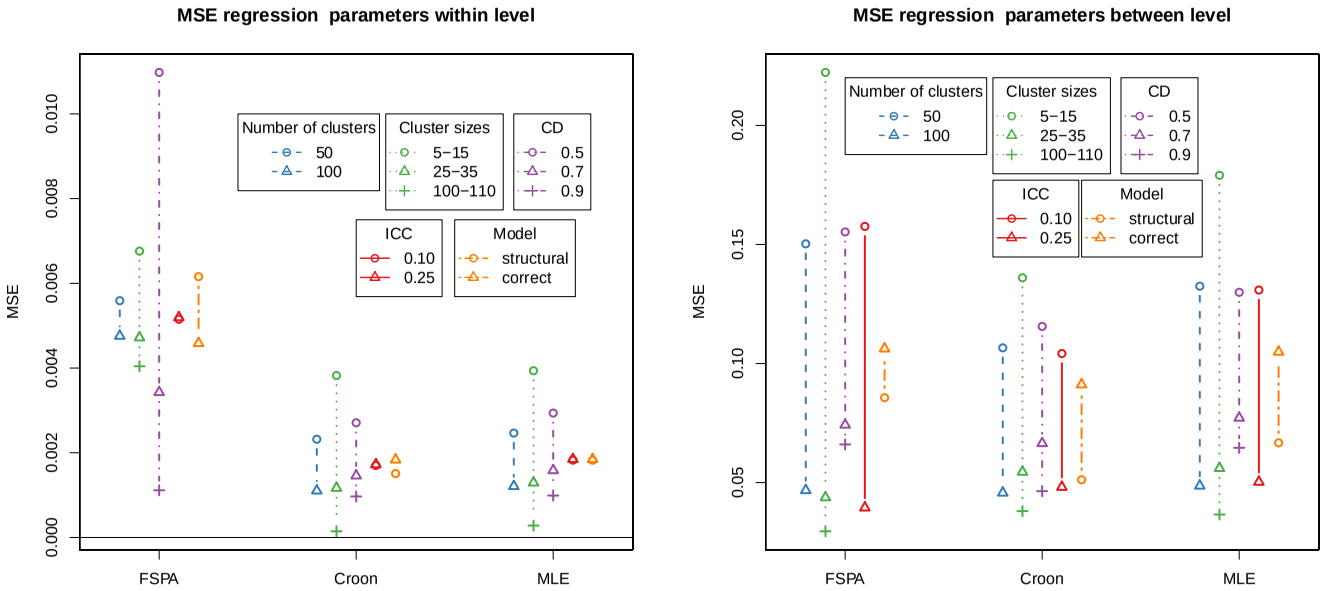
Coverage

The final performance criterium that we will discuss is the coverage.

Regression parameters The coverage of the Croon methods was the best out of the three methods. On both the within and between level, the coverage was always around 95%, as it should be. The coverage of MLE was sligthly below 95% in most settings, but was still acceptable in most settings. However, when there was a misspecification in the model, the coverage rate drops below 93% on the within level. Remember that these were parameters that should be unaffected by the misspecification.

Variance terms A similar pattern could be found for the coverage of the variance terms. The coverage of the within variance lied perfectly around 95% for the Croon method in all settings. The coverage of MLE dropped severely below 93 % when there was a structural misspecification. For the between level, the coverage of the Croon method also slightly dropped below 93%, but it was still much higher than the coverage rate of MLE.

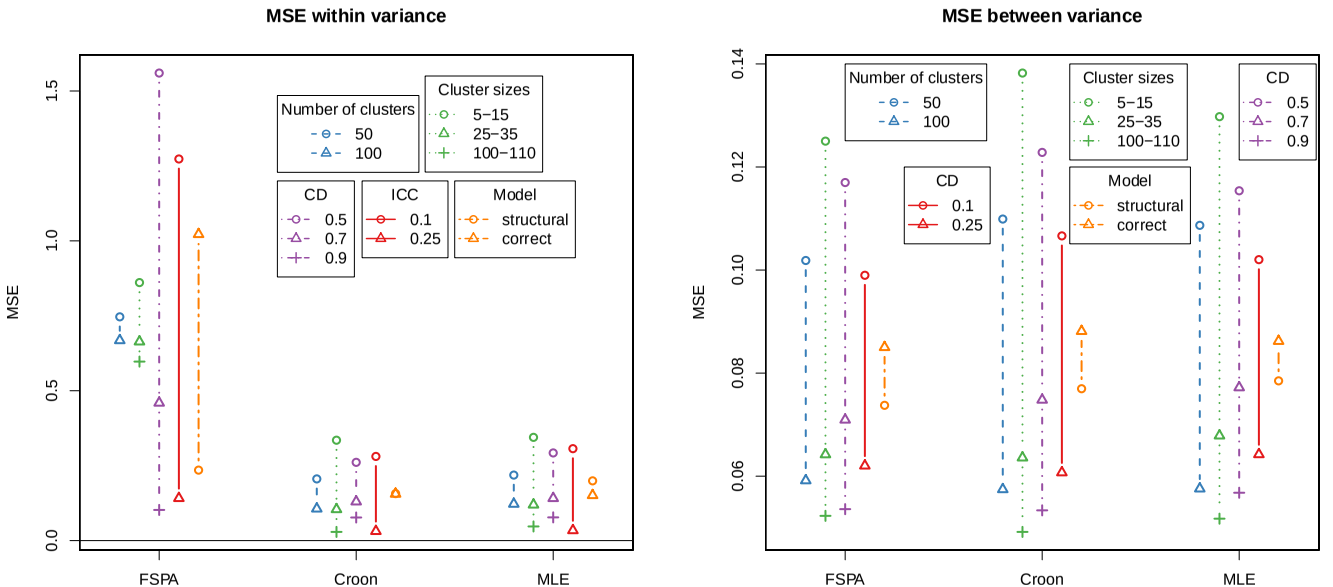
In conclusion, the method of Croon was a viable estimation method for multilevel SEM. The method of Croon performed just as well as MLE in almost all conditions. Firstly, it had fewer convergence



(a) MSE of the within regression parameters.

(b) MSE of the between regression parameters.

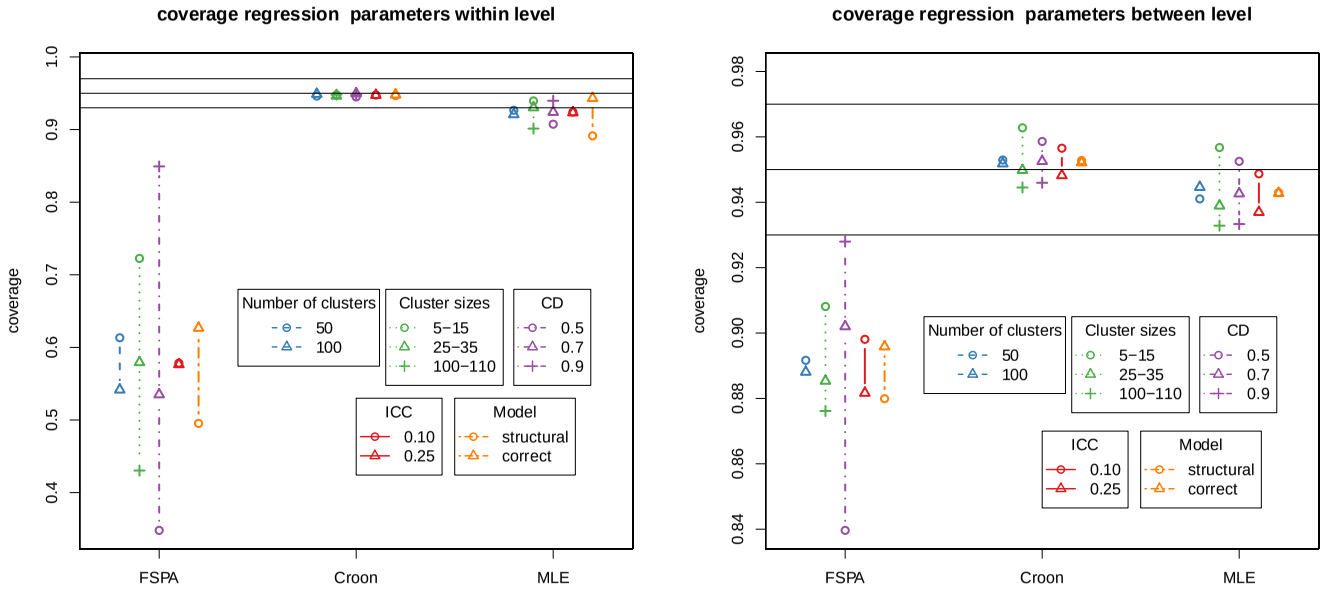
Figure 5.7 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the MSE of the regression parameters for the three methods; factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE).



(a) MSE of the within variance terms.

(b) MSE of the between variance terms.

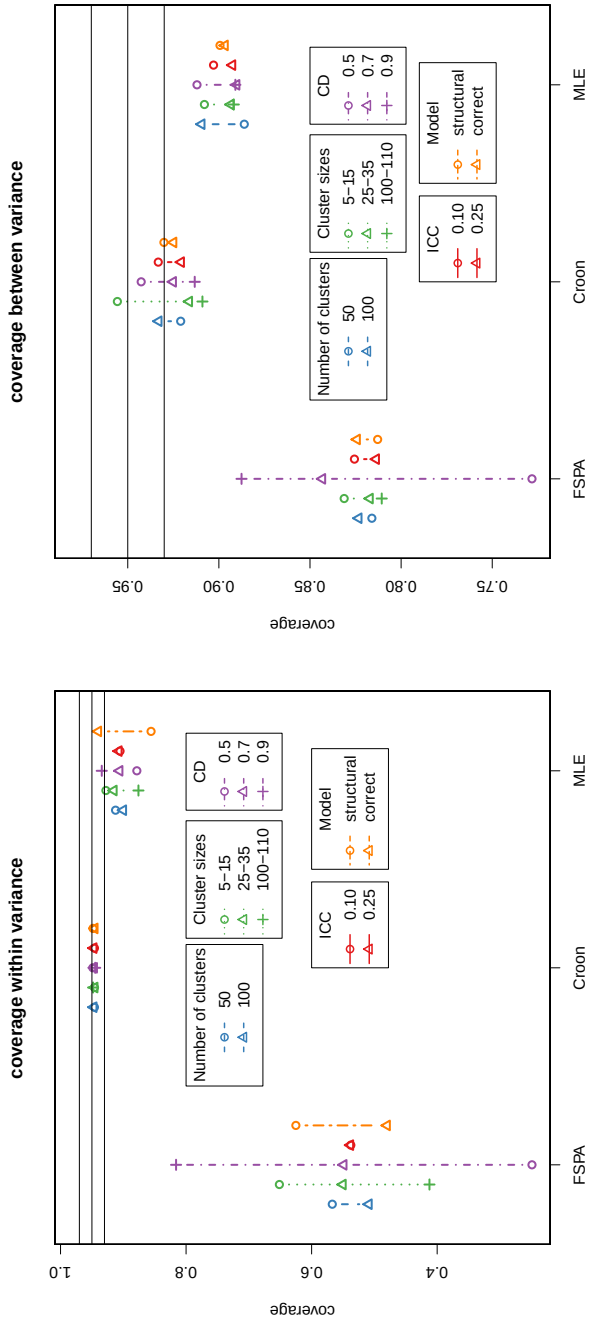
Figure 5.8 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the MSE of the variance terms for the three methods; factor score path analysis (FSPA), the Croon method (Croon), and maximum likelihood estimation (MLE).



(a) Coverage of the within regression parameters.

(b) Coverage of the between regression parameters.

Figure 5.9 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the coverage of the regression parameters for the three methods. The highest horizontal line represents a coverage of 97%, the lowest line a coverage of 93%, and the line in the middle is a coverage of 95%.



(a) Coverage of the variance terms.

(b) Coverage of the variance terms.

Figure 5.10 The influence of the number of clusters, the cluster sizes, the coefficient of determination (CD), the intra-class correlation (ICC) and the model specification on the coverage of the the variance terms for the three methods. The highest horizontal line represents a coverage of 97%, the lowest line a coverage of 93%, and the line in the middle is a coverage of 95%.

issues than MLE. It was less biased, both in the regression parameters as the variance terms. In most settings, the MSE was also lower, and the coverage rate was better in all settings. Our simulation study has also shown that the method of Croon can handle structural misspecifications better. The misspecification had less influence on the bias, MSE, and coverage rate of the other parameters in the model. In the next section, we illustrate the method of Croon with a real-world example.

5.6 Illustration

We applied the method of Croon to data from the 'Monitoring Information and Communications Technology (ICT) in Flemish Education 2012' (MICTIVO2). This is an educational policy- and practice-oriented scientific research project, set up to assess the impact of ICT at all levels of formal education (Goeman, Elen, Pynoo, & van Braak, 2015). The model for ICT-integration that was used, consists of four components, namely ICT-infrastructure and -policy, ICT-perception, ICT-compentencies, and ICT-use at the micro-level. Each component was measured by several variables. We only used the variables that are latent constructs, which were measured using Likert scales, ranging from 5-point to 7-point scales. Data were collected using several surveys, filled in by students, teachers, and principals. In total 723 principals, 2585 teachers, and 4887 pupils from 729 schools participated in the study. We only used a subset of the data, namely pupils from primary education and their principals. The pupils were only selected if their respective principal also participated in the study. This resulted in 2033 pupils and 47 principals, from 47 schools. We tested two models using MLE and the Croon method. For both estimation methods, the software package Mplus, version 8 was used (L. Muthén & Muthén, 2017).

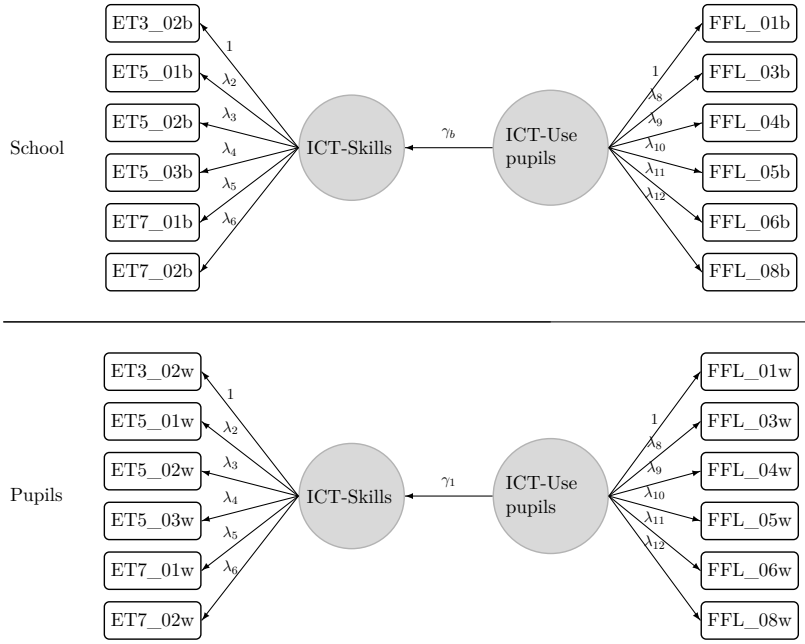


Figure 5.11 The first model used in the illustration.

5.6.1 Model 1

In the first model, we considered two latent variables, namely ICT-skills of the pupils and ICT-use by the pupils. Both variables were measured using 6 indicator items. We were interested in the relationship between both variables on the within and between level. The model is depicted in Figure 5.11. For the method of Croon, the procedure as described above was used. The results of this analysis, both the estimated parameters and their standard error, can be found in Table 5.6. For comparison reasons, the results when using SEM with MLE are also given. It can be seen that the results of both methods were very similar, including the standard errors.

5.6.2 Model 2

In the second model, we removed ICT-use by the pupils on the between level and also added two latent variables to the between

Table 5.6 The estimated parameters and their standard errors for the first illustration, both for the Croon method and MLE.

	Croon		MLE	
Parameter	Est.	(SE)	Est.	(SE)
Level 1 (pupils)				
Measurement models				
ICT-skills				
λ_1	1		1	
λ_2	1.010	0.026	1.013	0.026
λ_3	0.964	0.030	0.973	0.031
λ_4	0.901	0.030	0.909	0.030
λ_5	0.610	0.023	0.616	0.024
λ_6	0.564	0.022	0.569	0.022
ICT-use				
λ_7	1		1	
λ_8	0.722	0.042	0.741	0.042
λ_9	0.878	0.052	0.895	0.052
λ_{10}	0.887	0.049	0.895	0.049
λ_{11}	0.571	0.040	0.577	0.040
λ_{12}	0.786	0.065	0.844	0.066
Regressions				
γ_1	0.389	0.037	0.398	0.037
Variance				
ICT-skills	0.767	0.038	0.758	0.037
Level 2 (schools)				
Measurement model				
ICT-skills				
λ_1	1		1	
λ_2	1.010	0.026	1.013	0.026
λ_3	0.964	0.030	0.973	0.031
λ_4	0.901	0.030	0.909	0.030
λ_5	0.610	0.023	0.616	0.024
λ_6	0.564	0.022	0.569	0.022
ICT-use				
λ_7	1		1	
λ_8	0.722	0.042	0.741	0.042
λ_9	0.878	0.052	0.895	0.052
λ_{10}	0.887	0.049	0.895	0.049
λ_{11}	0.571	0.040	0.577	0.040
λ_{12}	0.786	0.065	0.844	0.066
Regressions				
γ_2	0.907	0.234	0.856	0.215
Variance				
ICT-skills	0.024	0.016	0.019	0.013

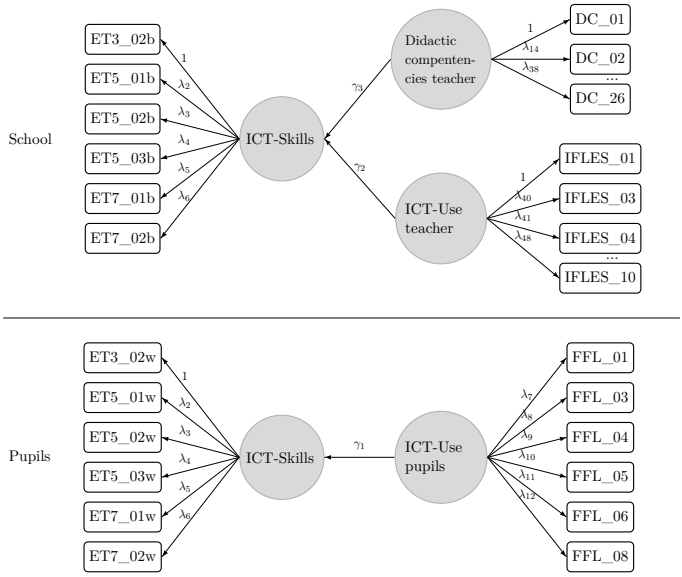


Figure 5.12 The second model used in the illustration.

level, namely the didactic competencies of the teacher and the ICT-use of the teacher during the class (see Figure 5.12). The former was measured by 26 items, while the latter was measured using 9 items. This model could not be estimated using MLE, due to convergence problems. However, the model could be estimated using the method of Croon. The results can be found in Table 5.7.

5.7 Discussion

In this paper, we proposed an alternative method to estimate multilevel structural equation models in a within-between framework. This method is an extension of the Croon method (Croon, 2002) for single-level factor score regression. A step-by-step procedure on how to perform the method was given. Then, a simulation study was used to assess the finite sample performance of the proposed method. For comparison purposes, naive multilevel factor score path analysis and maximum likelihood estimation were also

Table 5.7 The estimated parameters and their standard errors for the second illustration for the Croon method. The model could not be fitted for SEM using MLE.

	Croon	
Parameter	Est.	(SE)
Level 1 (pupils)		
Regressions		
γ_1	0.385	0.024
Variance		
ICT-skills	0.760	0.024
Level 2 (schools)		
Regressions		
ICT-skills		
γ_2	-0.036	0.070
γ_3	-0.015	0.091
Variance		
ICT-skills	0.039	0.013

included.

The simulation study showed that naive multilevel factor score path analysis does not perform well. All parameters that were considered showed bias, a high MSE, and a low coverage rate. This confirmed the results that were found in the single-level setting (Devlieger et al., 2016; Lu et al., 2011). When using the Croon method, the bias tends to disappear in the regression parameters, both at the within and the between level. At the within level, MLE also resulted in unbiased estimations. However, at the between level, bias was found in the regression parameters. This can be explained by the fact that the sample size is large enough at the within level, but the number of clusters was rather low at the between level. This confirmed previous results from the single level setting showing that the method of Croon can handle small sample sizes, here a small number of clusters, better than MLE (Devlieger & Rosseel, 2017; Lu et al., 2011). We also looked at the MSE, to evaluate if the Croon method is indeed less efficient than MLE, as could be expected from a step–wise method (Lüdtke et al., 2008; Wooldridge, 2002). We confirmed the results of Lüdtke et al. (2008) by finding that the MSE of the Croon method was larger than the MSE of MLE in the weakest condition, but similar or even lower in all other settings. The coverage rate of the Croon method was also similar or better than the coverage rate of MLE.

The simulation study also confirmed the results from the single level setting that the Croon method had fewer convergence issues than MLE. Note that this was also the case for settings where the model was correctly specified, so the non-convergence of MLE was not always an indication of misspecification. We also illustrated the convergence issues using an educational dataset, where MLE did not converge, while the Croon method experienced no problems. We consider this to be a major advantage of the Croon method over MLE. The number of successful replications could be increased for all three methods by using different estimators, changing the convergence criteria using different starting values or employing mul-

tiple starts. To avoid the covariance matrices being non-positive at the between level, stabilization techniques as suggested by Zitzmann (2018) could be used. However, in this paper, we chose to compare the basic methods.

In models that had misspecifications, it was found that a misspecification in the structural part had more influence on the regression parameters for MLE than for the method of Croon. No differences were found with regard to misspecifications in the measurement models. However, an advantage of step–wise methods over MLE, is that problems in the measurement model can be discovered more easily, since local fit indices for every measurement model can be obtained, while traditional SEM only gives global fit indices for the whole model.

Finally, some challenges still remain. Firstly, the procedure is rather complex and technical for an applied user to perform. However, the method is currently been implemented in the software package 'lavaan' (Rosseel, 2012). The fsr-function can be found in the development version of lavaan. Secondly, the Croon method can currently not handle random slopes. Future research should focus on developing a way to incorporate these random slopes. Thirdly, there are other alternatives for MLE that have been suggested in the past, such as the plausible values approach (Asparouhov & Muthén, 2010), Bayesian approaches discussed by Depaoli and Clifton (2015) and Zitzmann et al. (2016), and the method developed by Zitzmann (2018). Future research should compare all these methods to each other and to the method of Croon to determine which method is most suited for which settings.

References

Asparouhov, T., & Muthén, B. (2010). *Plausible values for latent variables using Mplus (Tech. Rep.).* (Tech. Rep.). Mplus Technical Report. doi: 10.1.1.310.3412

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014, jan). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, *22*(4), 520–540. doi: 10.1093/pan/mpu003

Bartlett, M. (1937, jul). The statistical conception of mental factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bollen, K. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109–121. doi: 10.1007/BF02296961

Bollen, K. (2018). Model Implied Instrumental Variables (MIIVs): An Alternative Orientation to Structural Equation Modeling. *Multivariate Behavioral Research*, *0*(0), 1–16. doi: 10.1080/00273171.2018.1483224

Chou, C.-P., Bentler, P. M., & Pentz, M. (2000). A two-stage approach to multilevel structural equation models: Application to longitudinal data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 33–49). Mahwah, NJ:: Lawrence Erblbaum Associates.

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007, mar). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. doi: 10.1037/1082-989X.12.1.45

Depaoli, S., & Clifton, J. P. (2015). A Bayesian Approach to Multi-

level Structural Equation Modeling With Continuous and Dichotomous Outcomes. *Structural Equation Modeling*, *22*(3), 327–351. doi: 10.1080/10705511.2014.937849

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi: 10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology*, *13*, 31–38. doi: 10.1027/1614-2241/a000130

Dijkstra, T. K. (2010). Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares. In V. E. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications (springer handbooks of computational statistics series, vol. ii)* (pp. 23–46). Heidelberg, Dordrecht, London, New York: Springer.

Goeman, K., Elen, J., Pynoo, B., & van Braak, J. (2015). Time for action! ICT Integration in Formal Education: Key Findings from a Region-wide Follow-up Monitor. *TechTrends*, *59*(5), 40–50. doi: 10.1007/s11528-015-0890-6

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*(2), 430–431. doi: 10.1093/biomet/74.2.430

Goldstein, H., & McDonald, R. P. (1988, dec). A general model for the analysis of multilevel data. *Psychometrika*, *53*(4), 455–467. doi: 10.1007/BF02294400

Holtmann, J., Koch, T., Lochner, K., Eid, M., Holtmann, J., Koch, T., ... Eid, M. (2016). A Comparison of ML , WLSMV , and Bayesian Methods for Multilevel Structural Equation Models in Small Samples : A Simulation Study A Comparison of ML , WLSMV , and Bayesian Methods for Multilevel Structural Equation Models in Small Samples : A Simulation S. *Multivariate Behavioral Research*, *51*(5), 661–680. doi: 10.1080/00273171.2016.1208074

Hoshino, T., & Bentler, P. M. (2013). Bias in Factor Score Regression and a Simple Solution. In A. R. de Leon & K. C. Chough (Eds.), *Analysis of mixed data- methods & applications* (pp. 43–61). Chapman and Hall. doi: 10.1201/b14571-5

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*(2), 157–170. doi: 10.1111/j.1467-9574.2009.00445.x

Hunter, J. E., & Gerbing, D. (1982). Unidimensional measurement, second order factor analysis, and causal models. *Research in Organizational Behavior*, *4*(267-320).

Lance, C. E., Cornwell, J. M., & Mulaik, S. A. (1988, apr). Limited Information Parameter Estimates for Latent or Mixed Manifest and Latent Variable Models. *Multivariate Behavioral Research*, *23*(2), 171–187. doi: 10.1207/s15327906mbr2302_3

Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*(4), 763–772. doi: 10.1093/biomet/77.4.763

Li, X., & Beretvas, S. N. (2013). Sample Size Limits for Estimating Upper Level Mediation Models Using Multilevel SEM. *Structural Equation Modeling*, *20*(2), 241–264. doi: 10.1080/10705511.2013.769391

Lu, I. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011, sep). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, *28*(3), 258–268. doi: 10.1016/j.ijresmar.2011.03.006

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011, dec). A 2 * 2 taxonomy of multilevel latent contextual models: Accuracy-bias trad-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467. doi: 10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-

Level Effects in Contextual Studies. *Psychological Methods*, *13*(3), 203–229. doi: 10.1037/a0012869

McDonald, R. P., & Goldstein, H. (1989, nov). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 215–232. doi: 10.1111/j.2044-8317.1989.tb00911.x

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*(1), 45–58. doi: 10.18148/srm/2009.v3i1.666

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585. doi: 10.1007/BF02296397

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological methods & research*, *22*, 376–398. doi: 10.1177/0049124194022003006

Muthén, B., & Asparouhov, T. (2009). Multilevel Regression Mixture Analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *172*(3), 639–657. doi: 10.1111/j.1467-985X.2009.00589.x

Muthén, L., & Muthén, B. (2017). *Mplus User's Guide.* Los Angeles, CA.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical computing. Retrieved from `https://www.r-project.org/`

Rabe-hesketh, S., Skrondal, A., & Andrew, P. (2004). Generalized Multilevel Structural Equation Modeling. *Psychological Methods*, *69*(2), 167–790. doi: 10.1007/BF02295939

Rasbash, J., Charlton, C., Browne, W., Healy, M., & Cameron, B. (2005). *MLwiN Version 2.02.* Centre for Multilevel Modelling, University of Bristol.

Rosseel, Y. (2012). lavaan : An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:

10.18637/jss.v048.i02

Schumacker, R., & Lomax, R. (1996). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Searle, S., Casella, G., & McCullouch, C. (1992). *Variance components*. New York: John Wiley and Sons. doi: 10.1002/9780470316856

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. doi: 10.1007/BF02296196

Takane, Y., & Hwang, H. (2018). Comparisons among several consistent estimators of structural equation models. *Behaviormetrika*, *45*(1), 157–188. doi: 10.1007/s41237-017-0045-5

Thomson, G. (1934). The meaning of i in the estimate of g. *British Journal of Psychology*, *25*, 92–99. doi: 10.1111/j.2044-8295.1934.tb00728.x

Thomson, G. (1938, feb). Methods of Estimating Mental Factors. *Nature*, *141*(3562), 246–246. doi: 10.1038/141246a0

Thurstone, L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press. doi: 10.1037/h0075959

Valluzzi, J. L., Larson, S. L., & Miller, G. E. (2003). Indications and Limitations of Structural Equation Modeling in Complex Surveys: Implications for an Application in the Medical Expenditure Panel Survey (MEPS). In *Joint statistical meetings - section on survey research methods indications* (pp. 4345–4352). doi: 10.1.1.493.467

van Erp, S., Mulder, J., & Oberski, D. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. doi: http://dx.doi.org/10.1037/met0000162

Venables, W., & Ripley, B. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. Retrieved from `http://www.stats.ox.ac.uk/pub/MASS4`

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press. doi: 10.1.1.464.9804

Zitzmann, S. (2018). A Computationally More Efficient and More Accurate Stepwise Approach for Correcting for Sampling Error and Measurement Error. *Multivariate Behavioral Research*, *0*(0), 1–21. doi: 10.1080/00273171.2018.1469086

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian Approach for Estimating Multilevel Latent Contextual Models. *Structural Equation Modeling*, *23*(5), 661–679. doi: 10.1080/10705511.2016.1207179

Appendices

5.A Variance of the between latent variable

In this appendix, we derive the variance of the between latent variable $\boldsymbol{\eta}_b$. First, we note that the observed cluster means $\mathbf{y}_b$ can also be written as:

$$\begin{aligned} \mathbf{y}_b &= \mathbf{y}_b^* + \frac{1}{n_j}\sum_i^{n_j} y_{wij}^* && (5.25)\\ &= \mathbf{y}_b^* + \bar{\mathbf{y}}_{wj}^*, && (5.26) \end{aligned}$$

To calculate the variance of $\boldsymbol{\eta}_b$, we use equation (5.18).

$$\begin{aligned} var(F_b) &= var(\mathbf{A}_b\mathbf{y}_b)\\ &= \mathbf{A}_b var(\mathbf{y}_b)\mathbf{A}_b^{'}\\ &= \mathbf{A}_b var(\mathbf{y}_b^* + \bar{\mathbf{y}}_{wj}^*)\mathbf{A}_b^{'}\\ &= \mathbf{A}_b [var(\mathbf{y}_b^*) + var(\bar{\mathbf{y}}_{wj}^*)]\mathbf{A}_b^{'}\\ &= \mathbf{A}_b [var(\mathbf{y}_b^*) + \frac{1}{n_j} var(\mathbf{y}_w^*)]\mathbf{A}_b^{'}\\ &= \mathbf{A}_b [var(\boldsymbol{\Lambda}_b\boldsymbol{\eta}_b + \boldsymbol{\epsilon}_b) + \frac{1}{n_j} var(\mathbf{y}_w^*)]\mathbf{A}_b^{'}\\ &= \mathbf{A}_b [var(\boldsymbol{\Lambda}_b\boldsymbol{\eta}_b) + var(\boldsymbol{\epsilon}_b) + \frac{1}{n_j} var(\mathbf{y}_w^*)]\mathbf{A}_b^{'}\\ &= \mathbf{A}_b [\boldsymbol{\Lambda}_b var(\boldsymbol{\eta}_b)\boldsymbol{\Lambda}_b^{'} + \boldsymbol{\Theta}_b + \frac{1}{n_j} var(\mathbf{y}_w^*)]\mathbf{A}_b^{'}\\ &= \mathbf{A}_b\boldsymbol{\Lambda}_b var(\boldsymbol{\eta}_b)\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'} + \mathbf{A}_b\boldsymbol{\Theta}_b\mathbf{A}_b^{'} + \frac{1}{n_j}\mathbf{A}_b var(\mathbf{y}_w^*)\mathbf{A}_b^{'} \end{aligned}$$

From this, we derive that

$$\begin{aligned} var(\boldsymbol{\eta}_b) =& (\mathbf{A}_b\boldsymbol{\Lambda}_b)^{-1}[var(\boldsymbol{F}_b) - \mathbf{A}_b\boldsymbol{\Theta}_b\mathbf{A}_b^{'}\\ &- \frac{1}{n_j}\mathbf{A}_b var(\mathbf{y}_w^*)\mathbf{A}_b^{'}](\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'})^{-1}, \end{aligned} \tag{5.27}$$

An estimate of $var(\boldsymbol{\eta}_b)$ can be found by replacing the population values of the right-hand side elements of equation (5.27) by its sample-based estimates. The cluster-specific n_j is replaced by $\tilde{n}$, defined as

$$\tilde{n} = \frac{N^2 - \Sigma_{j=1}^{J} n_j^2}{N(J-1)}$$

(Searle, Casella, & McCullouch, 1992).
Since $var(\boldsymbol{F}_b) = \mathbf{A}_b var(\mathbf{y}_b)\mathbf{A}_b^{'}$, we can rewrite this as:

$$\begin{aligned}
var(\boldsymbol{\eta}_b) &= (\mathbf{A}_b\boldsymbol{\Lambda}_b)^{-1}[\mathbf{A}_b var(\mathbf{y}_b)\mathbf{A}_b^{'} - \mathbf{A}_b\boldsymbol{\Theta}_b\mathbf{A}_b^{'} \\
&\quad -\frac{1}{n_j}\mathbf{A}_b var(\mathbf{y}_w^*)\mathbf{A}_b^{'}](\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'})^{-1} \\
&= (\mathbf{A}_b\boldsymbol{\Lambda}_b)^{-1}\mathbf{A}_b[var(\mathbf{y}_b) - \boldsymbol{\Theta}_b \\
&\quad -\frac{1}{n_j}var(\mathbf{y}_w^*)]\mathbf{A}_b^{'}(\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'})^{-1} \\
&= (\mathbf{A}_b\boldsymbol{\Lambda}_b)^{-1}\mathbf{A}_b[var(\mathbf{y}_b) - \frac{1}{n_j}var(\mathbf{y}_w^*) \\
&\quad -\boldsymbol{\Theta}_b]\mathbf{A}_b^{'}(\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'})^{-1} \\
&= (\mathbf{A}_b\boldsymbol{\Lambda}_b)^{-1}\mathbf{A}_b[var(\mathbf{y}_b^*) - \boldsymbol{\Theta}_b]\mathbf{A}_b^{'}(\boldsymbol{\Lambda}_b^{'}\mathbf{A}_b^{'})^{-1} \quad (5.28)
\end{aligned}$$

Therefore, there is no need to actually calculate the factor scores and their covariance matrix explicitly. All we need are the factor score matrices $\mathbf{A}_b$ and the parameters from the measurement blocks in $\boldsymbol{\Lambda}_b$ and $\boldsymbol{\Theta}_b$.

5.B Covariance between latent variables

In this appendix, we derive the covariance between two between latent variables $\boldsymbol{\eta}_{b1}$ and $\boldsymbol{\eta}_{b2}$.

$$\begin{aligned}
cov(F_{b1}, F_{b2}) &= cov(\mathbf{A}_{b1}\mathbf{y}_{b1}, \mathbf{A}_{b2}\mathbf{y}_{b2}) \\
&= \mathbf{A}_{b1} cov(\mathbf{y}^*_{b1}, \mathbf{y}^*_{b2}) \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} cov(\mathbf{y}^*_{b1} + \bar{\mathbf{y}}^*_{w1j}, \mathbf{y}^*_{b2} + \bar{\mathbf{y}}^*_{w2j}) \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} [cov(\mathbf{y}^*_{b1}, \mathbf{y}^*_{b2}) + cov(\mathbf{y}^*_{b1}, \bar{\mathbf{y}}^*_{w2j}) \\
&\quad + cov(\bar{\mathbf{y}}^*_{w1j}, \mathbf{y}^*_{b2}) + cov(\bar{\mathbf{y}}^*_{w1j}, \bar{\mathbf{y}}^*_{w2j})] \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} [cov(\mathbf{y}^*_{b1}, \mathbf{y}^*_{b2}) + \frac{1}{n_j} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2})] \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} [cov(\boldsymbol{\Lambda}_{b1}\boldsymbol{\eta}_{b1} + \boldsymbol{\epsilon}_{b1}, \boldsymbol{\Lambda}_{b2}\boldsymbol{\eta}_{b2} + \boldsymbol{\epsilon}_{b2}) \\
&\quad + \frac{1}{n_j} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2})] \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} [cov(\boldsymbol{\Lambda}_{b1}\boldsymbol{\eta}_{b1}, \boldsymbol{\Lambda}_{b2}\boldsymbol{\eta}_{b2}) + \frac{1}{n_j} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2})] \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} [\boldsymbol{\Lambda}_{b1} cov(\boldsymbol{\eta}_{b1}, \boldsymbol{\eta}_{b2}) \boldsymbol{\Lambda}^{'}_{b2} + \frac{1}{n_j} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2})] \mathbf{A}^{'}_{b2} \\
&= \mathbf{A}_{b1} \boldsymbol{\Lambda}_{b1} cov(\boldsymbol{\eta}_{b1}, \boldsymbol{\eta}_{b2}) \boldsymbol{\Lambda}^{'}_{b2} \mathbf{A}^{'}_{b2} \\
&\quad + \frac{1}{n_j} \mathbf{A}_{b1} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2}) \mathbf{A}^{'}_{b2}
\end{aligned}$$

From this, we derive that

$$\begin{aligned}
cov(\boldsymbol{\eta}_{b1}, \boldsymbol{\eta}_{b2}) =& (\mathbf{A}_{b1}\boldsymbol{\Lambda}_{b1})^{-1} [cov(F_{b1}, F_{b2}) \\
&- \frac{1}{n_j} \mathbf{A}_{b1} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2}) \mathbf{A}^{'}_{b2}] (\boldsymbol{\Lambda}^{'}_{b2} \mathbf{A}^{'}_{b2})^{-1},
\end{aligned} \tag{5.29}$$

An estimate of $cov(\boldsymbol{\eta}_{b1}, \boldsymbol{\eta}_{b2})$ can again be found by replacing the population values of the right-hand side elements of equation (5.29) by its sample-based estimates. The cluster-specific n_j is replaced by $\tilde{n}$.

Since $cov(F_{b1}, F_{b2}) = \mathbf{A}_{b1} cov(\mathbf{y}^*_{b1}, \mathbf{y}^*_{b2}) \mathbf{A}'_{b1}$, we can rewrite this as:

$$\begin{aligned} cov(\boldsymbol{\eta}_{b1}, \boldsymbol{\eta}_{b2}) &= (\mathbf{A}_{b1}\boldsymbol{\Lambda}_{b1})^{-1}[\mathbf{A}_{b1} cov(\mathbf{y}^*_{b1}, \mathbf{y}^*_{b2}) \mathbf{A}'_{b1} \\ &\quad - \frac{1}{n_j} \mathbf{A}_{b1} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2}) \mathbf{A}'_{b2}](\boldsymbol{\Lambda}'_{b2} \mathbf{A}'_{b2})^{-1} \\ &= (\mathbf{A}_{b1}\boldsymbol{\Lambda}_{b1})^{-1} \mathbf{A}_{b1}[cov(\mathbf{y}^*_{b1}, \mathbf{y}^*_{b2}) \\ &\quad - \frac{1}{n_j} cov(\mathbf{y}^*_{w1}, \mathbf{y}^*_{w2})] \mathbf{A}'_{b2} (\boldsymbol{\Lambda}'_{b2} \mathbf{A}'_{b2})^{-1} \end{aligned} \tag{5.30}$$

So, similar to the variance expression, there is no need to compute the factor scores and their covariances explicitly.

5.C Simulating raw data

step 4 can only be performed if the software accepts sample statistics as input. This is true for most SEM packages. However, regression based software (such as MLwiN (Rasbash et al., 2005) or the R-package 'lme4' (Bates et al., 2015)) only accepts raw data. These data can be generated by using $\hat{\boldsymbol{\Sigma}}_{\eta_w}$ and $\hat{\boldsymbol{\Sigma}}_{\eta_b}$ and the original factor scores. The within and between factor scores are transformed separately, each with their own method:

- The between factor scores $\mathbf{F}_b$ are transformed to have the corrected covariance matrix. This is done using the following formula:

  $$\mathbf{F}_{bt} = var(\mathbf{F}_b)^{-\frac{1}{2}} \; \mathbf{F}_b \; \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}_{\eta_b}. \tag{5.31}$$

  This formula rescales the variables, so that their covariance matrix corresponds to the corrected covariance matrix ($\hat{\boldsymbol{\Sigma}}_{\eta_b}$).

- For each cluster, the within factor scores $\mathbf{F}_{wsj}$ are simulated from a multivariate normal distribution with the means equal to zero and the covariance matrix equal to the estimated within covariance matrix. The simulated data $\mathbf{F}_{wsj}$ are then

transformed to fit the given parameters (means and covariance matrix) exactly using the following formula:

$$\mathbf{F}_{wtj} = [var(\mathbf{F}_{wsj})^{-\frac{1}{2}}\ \mathbf{F}_{wsj}\ \hat{\mathbf{\Sigma}}_{\eta_w}^{\frac{1}{2}}] - mean(\mathbf{F}_{wtj}). \quad (5.32)$$

- Finally, the between and within factor scores are combined by adding the cluster means $\mathbf{F}_{btj}$ to all the elements of the respective $\mathbf{F}_{wtj}$. The separate clusters are then stacked together, creating one big dataset.

Note that Formula 5.32 implies that every cluster size has to be at least as large as the number of variables in $\hat{\mathbf{\Sigma}}_{\eta_w}$, otherwise $var(\mathbf{F}_{wsj})^{-\frac{1}{2}}$ cannot be calculated. However, when the cluster size is too small to be able to transform the simulated data using formula 5.32, the original factor scores can be transformed using the following formula:

$$\mathbf{F}_{wtj} = \mathbf{S}_{Fw}^{-\frac{1}{2}}\ \mathbf{F}_{wj}\ \hat{\mathbf{\Sigma}}_{\eta_w}^{\frac{1}{2}}. \quad (5.33)$$

In contrast to $var(\mathbf{F}_{wsj})^{-\frac{1}{2}}$, $\mathbf{S}_{Fw}^{-\frac{1}{2}}$ is based on all the clusters and can thus be calculated. Note that this will not affect the estimates of the regression parameters or the estimate of the within variance. However, it will have a slight influence on the between variance, if many such small clusters occur in the data.

6 General discussion

At the beginning of this dissertation, I had three main objectives. In this discussion, I will review the progress that has been made and the challenges that remain for each of these objectives.

6.1 Comparing methods

My first objective was to compare stepwise methods to each other and to SEM, using MLE. In chapter 2, a simulation study showed that the bias correcting method is to be preferred over the bias avoiding method, because the latter still results in bias when the regression parameters are standardized. It was also not possible to extend the bias avoiding method to the mediational setting. Then, the best bias correcting method had to be chosen. In chapter 3, it is argued that the method of Croon is to be prefered over the method of Hoshino and Bentler, because the former is more general, in the sense that it works with any predictor and any estimator. The first conclusion of this dissertation was thus that the method of Croon is the best of the stepwise methods. This conclusion was also confirmed by Takane and Hwang (2018), who compared the bias avoiding method of Skrondal and Laake, the method of Croon and consistent partial least squares (PLSc). PLSc is a bias correcting method proposed by Dijkstra (2010).
After chosing the method of Croon as the best stepwise method, the focus was shifted to comparing this method to the 'gold standard' of simultaneously estimating all parameters using MLE. It was shown that in ideal circumstances (large sample sizes, normal data, correctly specified model), the method of Croon performs just

as well as MLE. In less ideal circumstances the method of Croon can even outperform MLE. Both in the single level (chapter 3) and the multilevel (chapter 5) setting, it was shown that the method of Croon handles small sample sizes and misspecifications better than MLE. The method more often results in succesful replications and shows less bias.

While the comparison of MLE and the method of Croon was quite extensive and spread out over all four empirical chapters, there are still some possibilities for future research. First of all, in all four studies, there was no missing data in the datasets. Further research is necessary to find out how the method of Croon should best handle missing data and how it's performance compares to MLE in the presence of missing data. Second, there are other alternatives to MLE, such as the instrumental variables approach Bollen (1996), the plausible values approach (Asparouhov & Muthén, 2010) and different Bayesian approaches. A large-scale study is needed to compare all these methods to the method of Croon and to MLE to determine which methods perform best in specific settings.

6.2 Expanding settings

The second objective of this dissertation was to expand the method of Croon to other settings. The method was successfully expanded to mediational settings in chapter 3 and to the within-between multilevel framework in chapter 5. These were important advances for the method of Croon, but there is still a lot of room for improvement, especially in the multilevel setting. In chapter 5, we limited ourself to the two-level within-between framework. The method should also be expanded to three-level settings (or more) and should be able to handle random slopes. One way to do this would be to use the generalized linear latent and mixed (GLLAMM)-framework developed by Rabe-Hesketh, Skrondal, and Zheng (2012). This framework extended multilevel regression models or generalized linear mixed models (GLMM) to include latent variables. The advan-

tage of this framework over the within-between framework is that it can handle several outcome types, including ordered and unordered categorical outcomes, and a wide range of models, including models with random slopes. If the method of Croon could somehow be combined with the GLLAMM-framework, the flexibility of the method would be greatly enhanced.
Another setting where the method of Croon does not work yet, is models where the matrix with the factor loadings is not of full column rank. If this matrix is not of full column rank, the correction factor of Croon ($\mathbf{A}_\xi \mathbf{\Lambda}_x \mathbf{\Lambda}_y^{'} \mathbf{A}_\eta^{'}$) will be singular and non-invertible. Since the inverse of this factor is needed, the correction of Croon can no longer be applied. One example of such a model is the Social Relations Model (SRM). In this model each item loads on one of the latent variables, called the family effect, but also loads on another more specific latent variable. This means that the family effect can be expressed in terms of all the other effects, resulting in a factor loading matrix that is not of full column rank. Loncke et al. (2018).

6.3 Performing hypothesis testing and model testing

The last objective of this dissertation was to also focus on statistical testing, such as drawing inference and model testing. In chapter 4, approximate fit indices and a model comparison test were developped, making it possible to both draw inference and evaluate the fit of the model. In chapter 5, an analytical standard error developped by Bakk, Oberski, and Vermunt (2014) was used. Both the model comparison test and analytical standard errors seemed to work well in the simulation studies. However, an inflated type I-error rate was found for the model comparison test when the sample size was low. Further research is needed to investigate how bad this inflation gets when the sample sizes become even smaller and more importantly on how this can be corrected. Existing corrections for small sample sizes, such as the Bartlett correction (Bartlett, 1937,

1954; Savalei, 2010) and the Swain correction (Swain, 1975) could be used.
In general, it was extensively tested how the point estimate of the regression parameter behaves in specific conditions, such as non-normality, small sample sizes and model misspecifications. It would be interesting to do the same for the standard errors and fit indices of the model.

6.4 Conclusion

In conclusion, the method of Croon was extended in several ways. It is now available in more settings, such as mediational settings and multilevel settings, can be used to draw inference, to perform model comparisons and to evaluate the global fit of the data. The method evolved from a way to get an accurate point estimate for a regression paramater to a full-fledged method to estimate structural equation models. This has some practical consequences for applied users. Models that can't be estimated using maximum likelihood, because there are more parameters to be estimated than the sample size or because the model simply does not converge, can now be estimated using the method of Croon. This implies that more information can be drawn from existing datasets, but also that the sample size requirements for new research projects can possibly be lowered, saving time and resources. In this way, I hope to have contributed to research in educational sciences specifically, but also to the broader social sciences. While great progress has been made regarding the method of Croon, there are still some opportunities left to further develop the method, which have been discussed above. I hope that the suggestions that have been made, lead to exciting new research in the future.

References

Asparouhov, T., & Muthén, B. (2010). *Plausible values for la-*

tent variables using Mplus (Tech. Rep.). (Tech. Rep.). Mplus Technical Report. doi: 10.1.1.310.3412

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014, jan). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, *22*(4), 520–540. doi: 10.1093/pan/mpu003

Bartlett, M. (1937, jul). The statistical conception of mental factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. doi: 10.1111/j.2044-8295.1937.tb00863.x

Bartlett, M. (1954). A Note on the Multiplying Factors for Various $\chi 2$ Approximations. *Journal of the Royal Statistical Society Series B (Methodological)*, *16*(2), 296–298.

Bollen, K. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109–121. doi: 10.1007/BF02296961

Dijkstra, T. K. (2010). Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares. In V. E. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (pp. 23–46). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-32827-8_2

Loncke, J., Eichelsheim, V. I., Branje, S. J., Buysse, A., Meeus, W. H., & Loeys, T. (2018). Factor score regression with social relations model components: A case study exploring antecedents and consequences of perceived support in families. *Frontiers in Psychology*, *9*(SEP), 1–19. doi: 10.3389/fpsyg.2018.01699

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel Structural Equation Modeling. In *Handbook of structural equation modeling* (pp. 512–531).

Savalei, V. (2010). Small Sample Statistics for Incomplete Nonnormal Data : Extensions of Complete Data Formulae and a Monte Carlo Comparison Small Sample Statistics for Incomplete Nonnormal Data : Extensions of Complete Data Formulae and a Monte Carlo Comparison. *Structural Equation*

Modeling, *17*(2), 241–264. doi: 10.1080/10705511003659375

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Unpublished doctoral dissertation). University of Adelaide, Department of Statistics.

Takane, Y., & Hwang, H. (2018). Comparisons among several consistent estimators of structural equation models. *Behaviormetrika*, *45*(1), 157–188. doi: 10.1007/s41237-017-0045-5

7 English summary

In educational research and other social sciences, the variables of interest are often latent, meaning they cannot be measured directly. Studying the relationships between these latent variables requires specialised techniques. In theory, the best technique is Structural Equation Modelling (SEM), using Maximum Likelihood Estimation (MLE). Using MLE, all parameters are estimated simultaneously, resulting in reliable and unbiased estimates of the regression parameters. While SEM is the 'gold standard' in theory, in reality it is not always very practical. The simultaneous estimation of all parameters requires large sample sizes. Smaller sample sizes lead to bias or even non-convergence, meaning that no solution is found. Therefore applied reseachers often use factor score regression instead. They first perform a FA for the separate latent variables, calculate factor scores and use these in a subsequent analysis, such as linear regression. However, this results in biased regression parameters. In general, there are two ways to eliminate this bias. A first strategy is to avoid the bias by using the Regression predictor for the independent variables and the Bartlett predictor for the dependent variables. A second strategy is to correct the bias, by correcting the variance-covariance matrix of the factor scores. In this dissertation, I tried to find out which strategy is best and then expanded this strategy.

In chapter 2, five methods to estimate the regression parameter between two latent variables were compared, namely naive factor score regression using the Regression predictor, naive factor score regression using the Bartlett predictor, factor score regression with bias avoiding, factor score regression with bias correcting and MLE.

For the bias correcting, the correcting formulas developed by Croon (2002) were used. The five methods were compared using analytical calculations and two simulation studies. The results showed that the two naive FSR techniques indeed resulted in biased regression parameters. The three other methods were unbiased, if the unstandardized parameterization was used. When the standardized parameterization was used, the bias avoiding method was also biased. This meant that only the bias correcting method of Croon could be a suitable alternative to SEM. It had a comparable bias, efficiency, MSE, power and type I-error rate as MLE.

It was already established that the method of Croon could perform equally wel as SEM, but in Chapter 3 it was examined if it could even outperform MLE. To do this, the method was first extended to the mediational setting. Then two simulation studies were set up. The first study showed that the method of Croon is more robust against misspecifications than MLE. The second study showed that the method of Croon can also handle small sample sizes better. MLE converges less and shows more bias in the regression parameters if there is a misspecification in the model or if the sample size is too small.

In Chapter 4, the focus was shifted from the point estimation of the regression parameters to the inference of the regression parameters and to the evaluation of the models. An approximated χ^2-test statistic for the method of Croon was introduced. This test statistic can be used to calculate several approximated fit indices, meaning that the global fit of the model can now be evaluated when using FSR. The test statistic can also be used to conduct model comparison tests. These new extensions were evaluated by comparing them to MLE in a simulation study and were illustrated using a real-world dataset.

Finally, in Chapter 5, the method of Croon was extended to the multilevel setting. In this study an analytical standard error that takes the stepwise nature of the procedure into account was used (Bakk, Oberski, & Vermunt, 2014). In a simulation study, the new

multilevel Croon method and its standard errors were compared to multilevel SEM using maximum likelihood estimation (MLE). The Croon method outperformed MLE with regard to convergence rate, bias, MSE, and coverage.

In conclusion, it was determined that the bias correcting method of Croon was the best way to eliminate the bias inherent to FSR. Then, this method was extended to mediational and multilevel settings, and fit indices and model comparison tests were developped. Using simulation studies, it was shown that the method of Croon performs at least equally well to MLE and even outperforms MLE in specific settings, such as small sample sizes and modelmisspecifications.

References

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014, jan). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, *22*(4), 520–540. doi: 10.1093/pan/mpu003

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

8 Nederlandse samenvatting

In de onderwijskunde en andere sociale wetenschappen zijn de onderzochte variabelen vaak latent. Dit wil zeggen dat ze niet rechtstreeks gemeten kunnen worden, zoals motivatie of intelligentie. Deze variabelen worden gemeten aan de hand van geobserveerde items, bijvoorbeeld door gebruik te maken van een vragenlijst. Het onderzoeken van verbanden tussen dergelijke latente variabelen vereist gespecialiseerde technieken, zoals structurele vergelijkingsmodellen (SEM), gebruik makend van een 'maximum likelhood' schatter (MLE). MLE schat alle parameters simultaan en onvertekend. Dit houdt in dat niet alleen de latente variabelen zelf geschat worden aan de hand van de geobserveerde items, maar ook meteen de verbanden tussen de latente variabelen. Theoretisch gezien is deze techniek de standaard, alleen is het in de realiteit niet altijd even praktisch. De simultane schatting van alle parameters vereist grote steekproefgroottes en een volledig correct gespecifieerd model. Kleinere steekproefgroottes leiden tot vertekende schatters of soms kan er zelf geen oplossing gevonden worden. Daarom gebruiken toegepaste onderzoekers vaak factorscore regressie (FSR) als alternatief. Eerst voeren ze een aparte factoranalyse (FA) uit om de latente variabelen zelf te meten. Op basis van de FA worden factorscores berekend, die een schatting zijn van de latente variabele. Ten slotte worden deze factorscores gebruikt in een vervolganalyse, zoals lineare regressie. Dit leidt echter tot vertekende regressieparameters. Er zijn twee manieren om deze vertekening te elimineren. Een eerste strategie probeert de vertekening te vermijden door twee verschillende predictoren te gebruiken voor de afhankelijke en onafhankelijke variabelen. De regressie predictor wordt gebruikt voor

de onafhankelijke variabele, terwijl de Bartlett predictor gebruikt wordt voor de afhankelijke variabele. De tweede strategie corrigeert de vertekening door de variantie-covariantie matrix van de factorscores te corrigeren. In deze thesis zocht ik uit welke strategie de beste is, en deze daarna heb ik deze uitgebreid.

In hoofdstuk 2 werden vijf methoden vergeleken om het verband tussen twee latente variabelen te schatten, namelijk naïeve factorscore regressie met een Regressie predictor, naïeve factorscore regressie met een Bartlett predictor, factorscore regressie die vertekening vermijdt, factorscore regressie die vertekening corrigeert en MLE. Voor het corrigeren van de vertekening werd gebruik gemaakt van de formules ontwikkeld door Croon (2002). De vijf methodes werden vergeleken aan de hand van analytische berekeningen en twee simulatiestudies. Hieruit bleek dat naïeve factorscore regressie inderdaad tot vertekende parameters leidde. De andere drie methodes waren onvertekend, zolang de parameters ongestandaardiseerd waren. Bij het standaardiseren van de parameters kon de vertekening niet langer vermeden worden, maar wel nog gecorrigeerd. De correctiemethode van Croon was dus het enige mogelijke alternatief voor MLE: de vertekening, efficiëntie, MSE, onderscheidingsvermogen en type I-fout waren gelijkaardig als die van SEM.

Nadat in hoofdstuk 2 vastgesteld werd dat de methode van Croon even goed kon presteren als MLE, werd in hoofdstuk 3 nagegaan of de methode zelfs beter kon zijn in bepaalde settings. Hiervoor werd de methode uitgebreid naar de mediatiesetting en voerden we twee simulatiestudies uit. De eerste studie toonde aan dat de methode van Croon beter om kan gaan met modelmisspecificaties dan MLE. Uit de tweede studie bleek dat hetzelfde geldt voor kleine steekproefgroottes. MLE bereikt minder vaak een oplossing en vertoont meer vertekening in de parameters als er een modelmisspecificatie is of als de steekproefgrootte te klein is.

In hoofstuk 4 werd de aandacht verlegd van de puntschatting van de parameters naar de inferentie van de regressieparameters en modelevaluatie. Een χ^2-teststatistiek voor de methode van Croon werd

ontwikkeld. Deze toetsingsgrootheid kan gebruikt worden om verschillende fitmaten te berekenen, waardoor de globale fit van een model nu geëvalueerd kan worden. De toetsingsgrootheid kan ook gebruikt om modelvergelijkingen uit te voeren. Deze nieuwe ontwikkelingen werden geëvalueerd door ze te vergelijken met MLE in een simulatiestudie en werden ook geïllustreerd aan de hand van een onderwijskundige dataset.
In hoofdstuk 5 werd de methode van Croon uitgebreid naar de multilevel setting. In deze studie werd ook voor de eerste keer een analytische standaardfout gebruikt die het stapsgewijze karakter van de procedure in rekening brengt (Bakk, Oberski, & Vermunt, 2014). Aan de hand van een simulatiestudie werden de nieuwe multilevel Croon methode en zijn standaardfouten vergeleken met multilevel SEM met MLE. De Croon methode presteerde beter dan MLE met betrekking tot convergentie, vertekening, MSE en coverage.
Kort samengevat werd vastgesteld dat de correctiemethode van Croon de beste manier is om de vertekening die inherent is aan FSR te elimineren. Daarna werd deze methode uitgebreid naar de mediatie en multilevel setting en werden fit indices en modelvergelijkingstoetsten ontwikkeld. Aan de hand van simulatiestudies werd getoond dat de methode van Croon minstens even goed en vaak zelfs beter dan MLE presteerde, vooral in settings met kleine steekproefgroottes en modelmisspecificaties.

Bibliografie

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014, jan). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, *22*(4), 520–540. doi: 10.1093/pan/mpu003

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah: Lawrence Erblbaum Associates, Inc.

9 Data Storage Fact Sheets

Data Storage Fact Sheets Chapter 2

```
% Name/identifier study: PhD dissertation Ines Devlieger, Chapter 2
% Author: Ines Devlieger                          % Date: 31/10/2019

1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Ines Devlieger
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: ines.devlieger@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Yves Rosseel
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: Yves.Rosseel@ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data
Management, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing
using factor score regression: A comparison of four methods.
Educational and Psychological Measurement, 76, 741--770.
doi: 10.1177/0013164415607618
```

```
* Which datasets in that publication does this sheet apply to?:
Scripts for data generation


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher?
[X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [] other (specify):

* Who has direct access to the raw data (i.e., without intervention
of another person)?
  - [X] main researcher
  - [X ] responsible ZAP
  - [X ] all members of the research group
  - [ ] all members of UGent
  - [] other (specify):


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: R scripts to generate
    the data, R scripts to analyze the generated data.
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and
how this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...
```

```
* On which platform are these other files stored?
  - [X] individual PC
  - [X ] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without
intervention of another person)?
  - [X] main researcher
  - [X ] responsible ZAP
  - [X ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:
```

Data Storage Fact Sheets Chapter 3

```
% Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Ines Devlieger, Chapter 3
% Author: Ines Devlieger
% Date: 07/11/2019


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Ines Devlieger
```

```
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: ines.devlieger@ugent.be

1b. Responsible Staff Member (ZAP)
----------------------------------------------------------
- name: Prof. dr. Yves Rosseel
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: Yves.Rosseel@ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data
Management, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
==========================================================
* Reference of the publication in which the datasets are reported:
Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis:
An alternative for SEM? Methodology, 13, 31--38.
doi: 10.1027/1614-2241/a000130

* Which datasets in that publication does this sheet apply to?:
Scripts for data generation.


3. Information about the files that have been stored
==========================================================


3a. Raw data
----------------------------------------------------------

* Have the raw data been stored by the main researcher?
 [X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [] other (specify):

* Who has direct access to the raw data (i.e., without intervention
```

```
of another person)?
  - [X] main researcher
  - [X ] responsible ZAP
  - [X ] all members of the research group
  - [ ] all members of UGent
  - [] other (specify):

3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: R scripts to generate
    the data, R scripts to analyze the generated data.
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and
how this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [X ] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without
intervention of another person)?
  - [X] main researcher
  - [X ] responsible ZAP
  - [X ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
```

```
   - address:
   - affiliation:
   - e-mail:
```

Data Storage Fact Sheets Chapter 4

```
% Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Ines Devlieger, Chapter 4
% Author: Ines Devlieger
% Date: 07/11/2019


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Ines Devlieger
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: ines.devlieger@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Yves Rosseel
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: Yves.Rosseel@ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data
Management, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Devlieger, I., Talloen, W.,& Rosseel, Y. (2019). New Developments
in Factor Score Regression: Fit Indices and a Model Comparison Test.
Educational and Psychological Measurement, 79(6), 1017--1037
doi: 10.1177/0013164419844552
```

```
* Which datasets in that publication does this sheet apply to?:
Scripts for data generation + raw data from illustration in paper


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher?
 [X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X ] research group file server
  - [X] other (specify): The raw data was provided by Johan
van Braak, Jan Elen en Katie Goeman, who conducted the original
study. As such, they also possess the raw data.

* Who has direct access to the raw data (i.e., without intervention
of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [X] other (specify): The raw data was provided by Johan
van Braak, Jan Elen en Katie Goeman, who conducted the original
study. As such, they also possess the raw data.


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported
results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: R scripts to generate
    the data, R scripts to analyze the raw and generated data.
```

```
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and
how this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without
intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:
```

Data Storage Fact Sheets Chapter 5

```
% Data Storage Fact Sheet

% Name/identifier study: PhD dissertation Ines Devlieger, Chapter 5
% Author: Ines Devlieger
% Date: 29/11/2019


1. Contact details
```

```
============================================================

1a. Main researcher
------------------------------------------------------------
- name: Ines Devlieger
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: ines.devlieger@ugent.be

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Prof. dr. Yves Rosseel
- address: H. Dunantlaan 2, 9000 Gent
- e-mail: Yves.Rosseel@ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data
Management, Faculty of Psychology and Educational Sciences,
Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Devlieger, I., & Rosseel, Y. (2019). Multilevel factor score
regression. Multivariate Behavioral Research,
doi: 10.1080/00273171.2019.1661817

* Which datasets in that publication does this sheet apply to?:
Scripts for data generation + raw data from illustration in paper


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher?
[X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
```

```
- [X] researcher PC
- [X] research group file server
- [X] other (specify): The raw data was provided by Johan
van Braak, Jan Elen en Katie Goeman, who conducted the original
study. As such, they also possess the raw data.

* Who has direct access to the raw data (i.e., without intervention
of another person)?
- [X] main researcher
- [X] responsible ZAP
- [X] all members of the research group
- [ ] all members of UGent
- [X] other (specify): The raw data was provided by Johan
van Braak, Jan Elen en Katie Goeman, who conducted the original
study. As such, they also possess the raw data.


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
- [ ] file(s) describing the transition from raw data to reported
results. Specify: ...
- [ ] file(s) containing processed data. Specify: ...
- [X] file(s) containing analyses. Specify: R scripts to generate
the data, R scripts to analyze the raw and generated data.
- [ ] files(s) containing information about informed consent
- [ ] a file specifying legal and ethical provisions
- [ ] file(s) that describe the content of the stored files and
how this content should be interpreted. Specify: ...
- [ ] other files. Specify: ...


* On which platform are these other files stored?
- [X] individual PC
- [X] research group file server
- [ ] other: ...

* Who has direct access to these other files (i.e., without
intervention of another person)?
- [X] main researcher
- [X] responsible ZAP
- [X] all members of the research group
```

```
- [ ] all members of UGent
- [ ] other (specify): ...


4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
- name:
- address:
- affiliation:
- e-mail:
```