Estimating links between latent variables using Structural Equation Modeling in R

by

Vera Plomp

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on wednesday August 12, 2020 at 15:30.

Student number: 4556224
Project duration: April 22, 2020 – August 12, 2020
Thesis committee: Dr. A. J. Cabo, TU Delft, supervisor
Prof. dr. ir. C. Vuik, TU Delft

.

TU Delft

Abstract

Structural equation modeling is a statistical analysis technique used to analyse structural relationships between observed variables and unobserved latent variables, and can be used to estimate links between latent variables. The technique is most commonly used in the field of psychology, and is not well known in the field of mathematics. Literature on structural equation modeling often lacks mathematical formulation. In this work, the mathematical theory of the method is discussed and where needed mathematical formulation is introduced. This is done by discussing the five steps in which the method can be summarized. After analysing the mathematical theory of the method, we illustrate the method by applying it to collected data, and we interpret the results.

Preface

This thesis has been written in order to obtain the degree of Bachelor of Science. The research has been conducted under supervision of Dr. A. J. Cabo on behalf of the department of Statistics of the faculty EEMCS at the University of Technology Delft.

I would like to thank my supervisor Dr. A. J. Cabo for her guidance during this project. Our weekly virtual meetings helped me stay on track with the project during these turbulent times, and I hope to meet her in person some day. I also would like to thank Prof. dr. ir. C. Vuik for taking a seat in my thesis committee.

Vera Plomp
Delft, July 2020

Contents

1 Introduction **1**

2 Structural Equation Modeling **3**
2.1 Model specification . . . 3
2.1.1 Set of equations . . . 3
2.1.2 Path diagrams . . . 5
2.1.3 Matrix notation . . . 6
2.2 Model identification . . . 6
2.2.1 Conditions for identification . . . 7
2.3 Model estimation . . . 7
2.3.1 Maximum likelihood fitting function . . . 7
2.3.2 Iterative procedure . . . 9
2.4 Model evaluation . . . 10
2.4.1 Measures of absolute fit . . . 11
2.4.2 Measures of relative fit . . . 11
2.5 Model modification . . . 12

3 PRIME research **15**
3.1 Research Plan . . . 15
3.1.1 Research question . . . 15
3.1.2 Hypotheses . . . 16
3.1.3 Data . . . 17
3.2 Formulating the model in R . . . 17
3.3 Fitting the Model in R . . . 19
3.3.1 Model identification and estimation . . . 19
3.3.2 Model evaluation and modification . . . 19
3.4 Results . . . 23
3.4.1 Interpretation of the estimates . . . 24

4 Conclusion **27**

5 Discussion **29**
5.1 Fit indices . . . 29
5.2 Result PRIME research . . . 29

A R-code **31**

Bibliography **33**

1 Introduction

Structural equation modeling (SEM) is a statistical analysis technique, which is especially used in the field of social sciences. It is used to analyse structural relationships between (unobserved) latent and (observed) measured variables. The ability of the technique to assign values to relationships between the unobserved latent constructs from observable variables is what makes it popular in social sciences. Motivation, for example, cannot be measured directly as one can measure length. Instead one can develop a hypothesis of motivation and use observed variables to measure it. Structural equation modeling can then be used to test the hypothesis using data of the observed variables.

However, structural equation modeling is not well known among mathematicians. This might have to do with the fact that criticism of the method often addresses pitfalls in the mathematical formulation. Since literature on SEM often lacks the mathematical formulation, this thesis will attempt to mathematically formulate the method of structural equation modeling. In addition, we will use SEM to analyse data collected in a course at TU Delft from PRIME (PRogramme of Innovation in Mathematics Education).

First we will discuss the steps involved in the procedure of structural equation modeling in chapter 2 and go into detail about each of the steps.

Following this, in chapter 3 we will discuss the PRIME research.

2

Structural Equation Modeling

Performing structural equation modeling can be summarized in five steps: model specification, model identification, model estimation, model evaluation and model modification. In this chapter we will take a closer look at these steps one by one.

2.1. Model specification

The first step in SEM is formulating the model to be tested. A general model consists of observed variables and latent variables. We have two types of latent variables: endogenous latent variables and exogenous latent variables. Endogenous latent variables are determined from within the model, meaning they are influenced by other latent variables of the model. Exogenous latent variables are not determined from within the model, the causes lie outside the model. Thus exogenous latent variables are not influenced by other latent variables in the model, but can influence other latent variables.
All the latent variables are unobserved, but they can be defined by the observed variables. Every latent variable is linked to a certain amount of observed variables. In SEM literature the coefficient that describes the relationship between the latent variable and an observed variable is called a factor loading, and we will also use this name. In model specification the observed variables that define the endogenous latent variables are separated from the observed variables that define the exogenous latent variables.
With these ingredients we can specify the model. There are several ways to specify a model for testing and estimation. We will discuss three formats: equations, path diagrams and matrix notation.

2.1.1. Set of equations

A general structural equation model can be divided in two types of equations: structural equations and measurement equations. The structural equations establish the relationships between the latent variables and the measurement equations define the latent variables by linking them to the observed variables. The general model can be written as:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \tag{2.1}$$

$$Y = \Lambda_y\eta + \epsilon \tag{2.2}$$

$$X = \Lambda_x\xi + \delta \tag{2.3}$$

Equation (2.1) is the structural equation, with η endogenous latent variables and ξ exogenous latent variables.
The matrices B, representing the effects of endogenous latent variables on each other, and Γ, representing the effects of exogenous latent variables on the endogenous latent variables, link the latent variables to each other. Finally, we have the error term ζ. It is assumed that ζ is uncorrelated with η and ξ.
Equations (2.2) and (2.3) are measurement equations, where Y and X are the observed variables defining the endogenous latent variables and the observed variables of the exogenous latent variables, respectively. These observed variables are linked to the endogenous and exogenous latent variables by

the matrices of factor loadings Λ_y and Λ_x. Finally, we also have error terms ϵ and δ. It is assumed that ϵ and δ are uncorrelated with each other and with η, ξ and ζ.

The model implied covariance matrix $\Sigma(\theta)$ is the covariance matrix implied by the parameters of the model and plays an important role in the estimation of the model parameters, as we will further discuss in section (2.3). $\Sigma(\theta)$ can be derived from the equations (2.1-2.3).

Proposition 2.1.1. *Let θ be the vector of parameters of the model. Then the model implied covariance matrix is given by*

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_Y(I-B)^{-1}(\Gamma\Phi\Gamma^\intercal + \Psi)[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal + \Theta_\epsilon & (\Lambda_X\Phi\Gamma^\intercal[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal)^\intercal \\ \Lambda_X\Phi\Gamma^\intercal[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal & \Lambda_X\Phi\Gamma_X^\intercal + \Theta_\delta \end{bmatrix}, \tag{2.4}$$

where Φ, Ψ, Θ_ϵ and Θ_δ are the covariance matrices of ξ, ζ, ϵ and δ, respectively.

Proof (see [5]). We can express $\Sigma(\theta)$ as functions of the parameters by calculating $\Sigma_{YY}(\theta)$, $\Sigma_{XY}(\theta)$ and $\Sigma_{XX}(\theta)$, since $\Sigma(\theta)$ can be expressed as:

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{YY}(\theta) & \Sigma_{YX}(\theta) \\ \Sigma_{XY}(\theta) & \Sigma_{XX}(\theta) \end{bmatrix} = \begin{bmatrix} \mathbb{E}(YY^\intercal) & \mathbb{E}(XY^\intercal)^\intercal \\ \mathbb{E}(XY^\intercal) & \mathbb{E}(XX^\intercal) \end{bmatrix} \tag{2.5}$$

We will start with the covariance matrix of Y.

$$\begin{aligned} \Sigma_{YY}(\theta) = \mathbb{E}(YY^\intercal) &= \mathbb{E}[(\Lambda_y\eta + \epsilon)(\Lambda_y\eta + \epsilon)^\intercal] \\ &= \mathbb{E}[(\Lambda_y\eta + \epsilon)(\eta^\intercal\Lambda_y^\intercal + \epsilon^\intercal)] \\ &= \mathbb{E}[\Lambda_y\eta\eta^\intercal\Lambda_y^\intercal] + \Theta_\epsilon \\ &= \Lambda_y\mathbb{E}[\eta\eta^\intercal]\Lambda_y^\intercal + \Theta_\epsilon \end{aligned} \tag{2.6}$$

where Θ_ϵ is the covariance matrix of the error term ϵ.

$$\begin{aligned} \eta\eta^\intercal &= [(I-B)^{-1}(\Gamma\xi + \zeta)][(I-B)^{-1}(\Gamma\xi + \zeta)]^\intercal \\ &= [(I-B)^{-1}(\Gamma\xi + \zeta)][(\Gamma\xi + \zeta)^\intercal[(I-B)^{-1}]^\intercal] \\ &= (I-B)^{-1}(\Gamma\Phi\Gamma^\intercal + \Psi)[(I-B)^{-1}]^\intercal, \end{aligned} \tag{2.7}$$

where Φ is the covariance matrix of ξ and Ψ is the covariance matrix of the residual ζ. Substituting equation (2.7) into equation (2.6) gives us

$$\mathbb{E}(YY^\intercal) = \Lambda_Y(I-B)^{-1}(\Gamma\Phi\Gamma^\intercal + \Psi)[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal + \Theta_\epsilon. \tag{2.8}$$

A similar calculation shows that the covariance matrix of X can be expressed as:

$$\begin{aligned} \Sigma_{XX}(\theta) = \mathbb{E}(XX^\intercal) &= \mathbb{E}[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)^\intercal] \\ &= \mathbb{E}[\Lambda_X\xi\xi^\intercal\Lambda_x^\intercal + \delta\delta^\intercal] \\ &= \Lambda_X\Phi\Lambda_X^\intercal + \Theta_\delta \end{aligned} \tag{2.9}$$

where Θ_δ is the covariance matrix of the error term δ. Finally, we can write the covariance matrix $\Sigma_{XY^\intercal}$ as:

$$\begin{aligned} \Sigma_{XY}(\theta) = \mathbb{E}(XY^\intercal) &= \mathbb{E}[(\Lambda_x\xi + \delta)(\Lambda_y\eta + \epsilon)^\intercal] \\ &= \mathbb{E}(\Lambda_X\xi\eta^\intercal) \\ &= \Lambda_x\mathbb{E}(\xi\eta^\intercal)\Lambda_Y^\intercal \\ &= \Lambda_X\mathbb{E}(\xi[(I-B)^{-1}(\Gamma\xi + \zeta)]^\intercal)\Lambda_Y^\intercal \\ &= \Lambda_X\Phi\Gamma^\intercal[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal \end{aligned} \tag{2.10}$$

Now we can combine the above results into the implied covariance matrix

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_Y(I-B)^{-1}(\Gamma\Phi\Gamma^\intercal + \Psi)[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal + \Theta_\epsilon & (\Lambda_X\Phi\Gamma^\intercal[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal)^\intercal \\ \Lambda_X\Phi\Gamma^\intercal[(I-B)^{-1}]^\intercal\Lambda_Y^\intercal & \Lambda_X\Phi\Gamma_X^\intercal + \Theta_\delta \end{bmatrix}, \tag{2.11}$$

where each element of the matrix is a function of model parameters. □

2.1.2. Path diagrams

Another way to describe a structural equation model is the most intuitive one. We can formulate the model by drawing a path diagram. Such a path diagram shows the hypothesized relationships among the variables and it can easily be translated to corresponding equations as described in section (2.1). In the path diagrams describing structural equation models, latent variables are generally presented by circles or ovals and the observed variables are presented by squares. Single headed arrows represent the hypothesized influence of one variable on another, where the variable that the head points to is the one who is being influenced. It can be seen as a directed relationship.
Double-headed arrows between two variables imply that they are related or associated. Double-headed arrows from a variable to itself represents the variance of the variable. Finally, if there are no arrows connecting two variables this means that no direct relationship has been hypothesized between those variables.

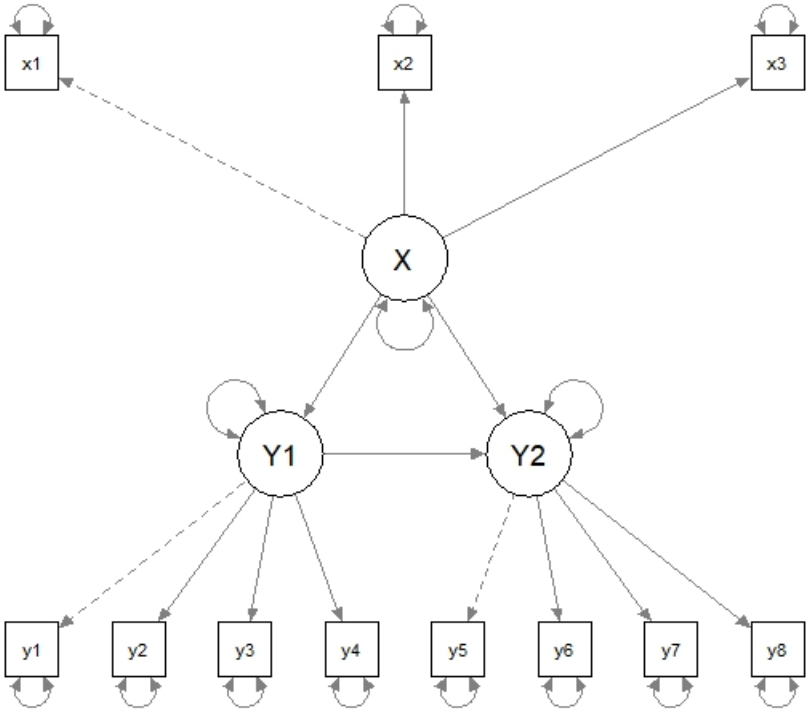


Figure 2.1: Path diagram of a structural equation model

An example of such a path diagram is shown in (2.1), where we can see that X, Y_1 and Y_2 are the latent variables. X is an exogenous latent variable, since it is not influenced by a latent variable, and it is measured by the observed variables $x_1, \cdots, x_3$. Y_1 and Y_2 are endogenous latent variables, since they are influenced by other latent variables, and are measured by the observed variables $y_1, \cdots, y_4$ and $y_5, \cdots, y_8$, respectively. By looking at the arrows we can see that X is hypothesized to influence Y_1 and Y_2, and Y_1 is hypothesized to influence Y_2.
Another thing that stands out are the dotted lines between the latent variables and one of the indicators, which means that the factor loading that links them is fixed. This has to do with identification of the model and will be further discussed in section (2.2).

2.1.3. Matrix notation

Based on the equations (2.1)-(2.3) from section (2.1) a structural equation model can be specified in matrix notation. The first equation can be expressed as:

$$\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix}=\begin{bmatrix}0 & \beta_{12} & \cdots & \beta_{1n}\\ \beta_{21} & 0 & & \beta_{2n}\\ \vdots & & \ddots & \vdots\\ \beta_{n1} & \beta_{n2} & \cdots & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix}+\begin{bmatrix}\gamma_{11} & \cdots & \gamma_{1n}\\ \vdots & \ddots & \vdots\\ \gamma_{n1} & \cdots & \gamma_{nn}\end{bmatrix}\begin{bmatrix}\xi_1\\ \vdots\\ \xi_n\end{bmatrix}+\begin{bmatrix}\zeta_1\\ \vdots\\ \zeta_n\end{bmatrix}$$

The first measurement equation can be written as

$$\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}=\begin{bmatrix}\lambda_{y11} & \cdots & \lambda_{y1n}\\ \vdots & \ddots & \vdots\\ \lambda_{yn1} & \cdots & \lambda_{ynn}\end{bmatrix}\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix}+\begin{bmatrix}\epsilon_1\\ \vdots\\ \epsilon_n\end{bmatrix}$$

and the second measurement equation as

$$\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}=\begin{bmatrix}\lambda_{x11} & \cdots & \lambda_{x1n}\\ \vdots & \ddots & \vdots\\ \lambda_{xn1} & \cdots & \lambda_{xnn}\end{bmatrix}\begin{bmatrix}\xi_1\\ \vdots\\ \xi_n\end{bmatrix}+\begin{bmatrix}\delta_1\\ \vdots\\ \delta_n\end{bmatrix}.$$

To illustrate this we can describe the path diagram (2.1) of the previous section in matrix notation. Suppose γ_{11} and γ_{21} are the effects of the exogenous variable X on Y_1 and Y_2, respectively. And suppose β_{12} specifies the effect of Y_1 on Y_2. Then the structural part of the model can be expressed as

$$\begin{bmatrix}Y_1\\ Y_2\end{bmatrix}=\begin{bmatrix}0 & \beta_{12}\\ 0 & 0\end{bmatrix}\begin{bmatrix}Y_1\\ Y_2\end{bmatrix}+\begin{bmatrix}\gamma_{11}\\ \gamma_{21}\end{bmatrix}X+\begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix}.$$

The first measurement equation can be written as

$$\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\end{bmatrix}=\begin{bmatrix}1 & 0\\ \lambda_{y21} & 0\\ \lambda_{y31} & 0\\ \lambda_{y41} & 0\\ 0 & 1\\ 0 & \lambda_{y62}\\ 0 & \lambda_{y72}\\ 0 & \lambda_{y82}\end{bmatrix}\begin{bmatrix}Y_1\\ Y_2\end{bmatrix}+\begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\\ \epsilon_5\\ \epsilon_6\\ \epsilon_7\\ \epsilon_8\end{bmatrix}$$

and the second measurement equation as

$$\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}1\\ \lambda_{x21}\\ \lambda_{x31}\end{bmatrix}X+\begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\end{bmatrix}.$$

2.2. Model identification

Before we can estimate the unknown parameters of the model, we have to establish whether it is possible to obtain consistent estimates of the unknown paramaters. This consideration is called model identification. We will now give the definition of identification.

Definition 2.2.1. *Let θ be the vector of parameters of the model and let $\Sigma(\theta)$ be the model implied covariance matrix. Now consider two vectors θ_1 and θ_2, which contain values for the parameters in θ. We say the model is identified if for all θ_1 and θ_2 satisfying $\Sigma(\theta_1)=\Sigma(\theta_2)$, we have $\theta_1=\theta_2$.*

In words, a model is said to be identified if there exists a unique solution for all of the model's parameters.
Parameters can also be over-identified, which means there is more than one way to estimate the parameter. When all the parameters are identified and one or more parameter is over-identified, the model is over-identified. Over-identified models are the models of interest in structural equation modeling.

2.2.1. Conditions for identification

We can say that the question of identification is whether the amount of unknown information in the model is less than or equal to or bigger than the amount of known information. The number of known variables equals the nonredundant elements in the population covariance matrix Σ, which is the the matrix of covariances between the observed variables. Since we know the number of observed variables, which we denote by p, the number of known values can be calculated by $\frac{1}{2}p(p+1)$.
The first necessary condition of identification is a test called the t-rule [8]. It uses the number of known variables to set a condition on how many unknown parameters the model can have in order to be identified. The t-rule states that the number of known values must be greater or equal to the number of unknown parameters in θ:

$$t \leq \frac{1}{2}(p)(p+1) \tag{2.12}$$

where the right-hand side of (2.12) is the number of nonredundant elements in the population covariance matrix Σ, and t is the number of free parameters in θ. If there are no correlated errors in the model, the number of free parameters can be calculated by $2 \cdot p + z$ [8], where z is the number of relationships between latent variables.
Thus we have $\frac{1}{2}(p)(p+1)$ equations in t unknowns and if the number of unknowns exceeds the number of equations, identification is not possible. This difference between known and unkown information equals the model's degrees of freedom (df), therefore the t-rule corresponds with having a non-negative df. To illustrate the t-rule we can use the model defined in section (2.1.2) of the path diagram (2.1). We can see that the model has eleven observed variables and there are three relationships between latent variables. Hence the t-rule holds: $t = 25 \leq \frac{1}{2}(p)(p+1) = 66$.

Another condition is that we need to establish a measurement scale for the latent variables in the model, since they are unmeasured and therefore do not have a scale. Obtaining consistent estimates is not possible if the latent variables do not have a scale. Every latent variable is measured by different observed variables, from which we can constrain one of the factor loadings that links them to the latent variables to one in order to set the scale for the latent variable. For every latent variable, one factor loading has to be fixed in order to be identified.

2.3. Model estimation

The basic SEM hypothesis equals

$$\Sigma = \Sigma(\theta), \tag{2.13}$$

where Σ is the population covariance matrix and $\Sigma(\theta)$ is the model implied covariance matrix (2.11) with θ containing the parameters of the model. Therefore the model parameters are estimated by minimizing the difference

$$(\Sigma - \Sigma(\theta)). \tag{2.14}$$

But since we do not know both Σ and $\Sigma(\theta)$, we actually minimize the difference

$$(S - \Sigma(\hat{\theta})), \tag{2.15}$$

where S is the sample covariance matrix and $\Sigma(\hat{\theta})$ is the model estimated covariance matrix with $\hat{\theta}$ containing the estimated parameters of the model. In words, we want those values for θ that minimize the difference between what we observe in the data and what the model implies.

2.3.1. Maximum likelihood fitting function

To minimize the difference between the sample covariance matrix and the model estimated covariance matrix $(S - \Sigma(\hat{\theta}))$ a help function, called a fitting function, is needed. The main desired property of such a function is that it can tell us how small the difference between the two covariance matrices is, and thus how ”close” our estimates are. We will now give the definition of a fitting function.

Definition 2.3.1. *A function $F : \mathbb{R}^n \to \mathbb{R}$ is called a fitting function if*

a) F *is continuous,*

b) $F \geq 0$

c) $F = 0 \iff S = \Sigma(\hat{\theta})$

The most commonly used fitting function in SEM is the maximum likelihood (ML) function.

Proposition 2.3.1. *The maximum likelihood fitting function $F_{ML}(\hat{\theta})$ is given by*

$$F_{ML}(\hat{\theta}) = ln|\Sigma(\hat{\theta})| + tr(S\Sigma(\hat{\theta})^{-1}) - ln|S| - p, \tag{2.16}$$

where p is the number of observed variables.

The derivation of the maximum likelihood fitting function is based on the assumption that Y and X, as defined in (2.2)-(2.3), follow a multivariate normal distribution and as a result of that, that S follows the Wishart distribution [10]. But even when Y and X violate the multivariate normality assumption, the maximum likelihood estimator has the desirable properties under less restrictive assumptions [8].
We will now give the proof of proposition (2.3.1).

Proof (see [5]). When we select a random sample from the multivariate normal population, the likelihood of finding a sample with covariance matrix S is given by the Wishart distribution

$$W(S, \Sigma, n) = \exp -\frac{1}{2}n \cdot tr(S\Sigma^{-1})|\Sigma|^{-\frac{1}{2}n}C, \tag{2.17}$$

where $n = N-1$, with N the sample size, and C contains all the constant terms of the Wishart distribution. When the structural equation model is a perfect fit, $S = \Sigma(\hat{\theta})$. Thus the likelihood ratio of the specified model to that of the perfect model is equal to

$$\begin{aligned} LR &= \frac{\exp\left(-\frac{1}{2}n \cdot tr(S\Sigma(\hat{\theta})^{-1})\right)|\Sigma(\hat{\theta})|^{-\frac{1}{2}n}C}{\exp\left(-\frac{1}{2}n \cdot tr(SS^{-1})\right)|S|^{-\frac{1}{2}n}C} \\ &= \exp\left(-\frac{1}{2}n \cdot tr(S\Sigma(\hat{\theta})^{-1})\right)|\Sigma(\hat{\theta})|^{-\frac{1}{2}n} \exp\left(\frac{1}{2}n \cdot tr(SS^{-1})\right)|S|^{\frac{1}{2}n}. \end{aligned} \tag{2.18}$$

Now we can take the natural logarithm of equation (2.18) to obtain

$$\begin{aligned} ln(LR) &= -\frac{1}{2}n\Big[tr(S\Sigma(\hat{\theta})^{-1}) + ln|\Sigma(\hat{\theta})| - tr(SS^{-1}) - ln|S|\Big] \\ &= -\frac{1}{2}n\Big[tr(S\Sigma(\hat{\theta})^{-1}) + ln|\Sigma(\hat{\theta})| - p - ln|S|\Big], \end{aligned} \tag{2.19}$$

where p is the number of observed variables involved in the model. Maximizing equation (2.19) is equivalent to minimizing

$$F_{ML}(\hat{\theta}) = ln|\Sigma(\hat{\theta})| + tr(S\Sigma(\hat{\theta})^{-1}) - ln|S| - p, \tag{2.20}$$

which we call the maximum likelihood fitting function. The maximum likelihood method requires S to be positive definite. □

We can see in equation (2.20) that when $\Sigma(\hat{\theta})$ equals S the value of the function will be zero, since the natural logarithms will cancel out and the trace of the identity matrix will equal p, the number of observed variables involved in the model. Hence, we want to estimate the model parameters such that the maximum likelihood function is minimized.

2.3.2. Iterative procedure

Solving for the unknown parameters θ relies on an iterative process. Starting values $\hat{\theta}$ are set for the parameter values and used to start the iterative process. We will show the starting values that the R-package 'lavaan', that we will be using in chapter (3), uses to start the iterative process. First we will discuss the starting values for the factor loadings, which are computed by using 'fabin 3' or also called two-stage least squares estimates.

Theorem 2.3.1. *Let λ be a factor loading. Let S_{ij} denote a submatrix of the sample covariance matrix and let s_{ij} denote a row vector of covariances between i and j. Then an estimate of λ is given by:*

$$\hat{\lambda}^{\intercal} = (S_{13}S_{33}S_{31})^{-1}S_{13}S_{33}^{-1}s_{32} \tag{2.21}$$

Proof (see [2]). Consider the measurement equation

$$Y = \Lambda g + \epsilon, \tag{2.22}$$

where $Y(n \times 1)$ is a vector of observed variables, $\Lambda(n \times m)$ a matrix of factor loadings, $g(m \times 1)$ a vector of latent variables and $\epsilon(n \times 1)$ a vector of residuals. We assume that all the variables have zero means and that $n \geq 3m$. Now we rewrite (2.22) as

$$\begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \mathrm{y}_3 \end{bmatrix} = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \\ \Lambda_3 \end{bmatrix} \mathrm{g} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}, \tag{2.23}$$

where $y_1(m \times 1)$, $y_2(p \times 1)$, $y_3(q \times 1)$, $\Lambda_1(m \times m)$, $\Lambda_2(p \times p)$, $\Lambda_3(q \times q)$, $\epsilon_1(m \times 1)$, $\epsilon_2(p \times 1)$, $\epsilon_3(q \times 1)$ and $\mathrm{f}(m \times 1)$. Furthermore, we assume that $p \geq m$ and $q \geq m$ and we set $\Lambda_1 = I$ for the sake of identification. Solving the first equation of (2.23) for f and substituting it in the second equation gives us

$$\mathrm{y}_2 = \Lambda_2 \mathrm{y}_1 + \epsilon_2 - \Lambda_2 \epsilon_1 \tag{2.24}$$

$$\epsilon_2 - \Lambda_2 \epsilon_1 = \mathrm{y}_2 - \Lambda_2 \mathrm{y}_1 \tag{2.25}$$

Since y_3 is uncorrelated with ϵ_1 and ϵ_2, it is also uncorrelated with the residual (2.25). Now the following equation holds

$$\mathbb{E}[\mathrm{y}_3(\mathrm{y}_2 - \Lambda_2 \mathrm{y}_1)^{\intercal}] = 0, \tag{2.26}$$

which is true if and only if

$$\Lambda_2 = \Sigma_{23}\Sigma_{31}(\Sigma_{13}\Sigma_{31})^{-1}, \tag{2.27}$$

where Σ_{23} and Σ_{31} are submatrices of the population covariance matrix. For the proof of the above used implication we refer you to [1]. Since the equation (2.23) is symmetric in y_2 and y_3 we also have

$$\Lambda_3 = \Sigma_{32}\Sigma_{21}(\Sigma_{12}\Sigma_{21})^{-1}. \tag{2.28}$$

The above results are estimation formulas if we use the corresponding sample covariance matrices instead of the population covariance matrices. But from these estimators better estimates can be derived. Suppose $q = 1$ in (2.23), thus y_2 only contains 1 variable. Then Λ_2 has only one row $\lambda_2(1 \times m)$. The estimate of Λ_2 can be obtained from (2.27)

$$\hat{\lambda_2} = s_{23}S_{31}(S_{13}S_{31})^{-1}, \tag{2.29}$$

where s_{23} is a row vector of covariances between the variable in y_2 and all the variables of y_3. By repeatedly selecting one variable to be in y_2 and the remaining variables to be in y_3, estimates of the rows of Λ are obtained. We will now show that (2.29) can be justified by the method of least squares. For the estimator (2.29), equation (2.24) becomes

$$y_2 = \lambda_2 y_1 + \epsilon_2 - \lambda_2 \epsilon_1, \tag{2.30}$$

which for $\lambda = \lambda_2$ and $u = \epsilon_2 - \lambda_2 \epsilon_1$ can be written as

$$y_2 = \lambda y_1 + u. \tag{2.31}$$

Multiplying by $y_3^\intercal$ and taking the expectation under the assumption that the variables are measured with zero means gives us

$$\begin{aligned} \mathbb{E}(y_2 y_3^\intercal) &= \mathbb{E}(\lambda y_1 y_3^\intercal) + \mathbb{E}(u y_3^\intercal) \\ \sigma_{23} &= \lambda \Sigma_{13} \end{aligned} \tag{2.32}$$

Replacing the population covariances with the corresponding sample covariances will lead to the relation

$$s_{23} = \lambda S_{13} + s_{u3}, \tag{2.33}$$

with the transpose

$$s_{32} = S_{31}\lambda^\intercal + s_{3u}. \tag{2.34}$$

Now we should use weighted least squares, because the covariance matrix of s_{3u} is not of the form $\sigma^2 I$ [2]. Thus we minimize $s_{u3}\Sigma_{33}^{-1}s_{3u}$ and for large samples we replace Σ_{33}^{-1} by S_{33}^{-1}. Applying this to (2.34) gives us the result:

$$\hat{\lambda^\intercal} = (S_{13}S_{33}S_{31})^{-1}S_{13}S_{33}^{-1}s_{32} \tag{2.35}$$

□

The other starting values are determined easily. For the residual variances of observed variables, the starting value is the total variance divided by two. For the residual variances of the latent variables, the starting values are set to 0.05. Finally, everything else has a starting value of 0.

The starting values can be substituted into the model implied covariance matrix. Minimizing the difference between the model implied covariance matrix and the sample covariance matrix, using the fitting function, will give us new parameter estimates. These new parameters will be substituted again and this process will continue until changes in parameter values have become acceptably small. To determine when the changes have become acceptably small, the relative and absolute function convergence tolerance are used.

Definition 2.3.2. *Let f_i be the value of the fitting function at the i-th iteration step. Then the relative function convergence tolerance is given by*

$$\frac{|f_i - f_{i-1}|}{min(|f_i|, |f_{i-1}|)}, \tag{2.36}$$

At every iteration i, we check whether the relative tolerance is smaller than or equal to a beforehand defined convergence criterion, which in the case of the R-package 'lavaan' is $1 \cdot 10^{-10}$. If this is the case, the minimization stops. But if $f(i)$ becomes near to zero, the relative tolerance grows to infinity. Therefore we also use the absolute tolerance.

Definition 2.3.3. *Let f_i be the value of the fitting function at the i-th iteration step. Then the absolute function convergence tolerance is given by*

$$|f_i - f_{i-1}|. \tag{2.37}$$

So at every iteration we also check if the absolute tolerance is smaller than or equal to a convergence criterion, which in 'lavaan' is set to $2.22 \cdot 10^{-15}$.
If either one of the convergence criteria is met, the minimization stops and we turn to the next step of structural equation modeling.

2.4. Model evaluation

Before we take a look at the estimates, we need to check if the model fits the data well. Because if the model does not fit the data well, the estimates are not reliable. Evaluating the model fit comes down to assessing the extent to which the observed sample covariance matrix S and the model estimated matrix $\Sigma(\hat{\theta})$ differ, and thus assessing the value of the maximum likelihood fitting function, shown in (2.3.1). To that end, a lot of model fit indices are available. We will discuss the most common model fit indices in structural equation modeling.
We will start by discussing a measure of overall fit, which will be used to define the other fit indices discussed in this section.

Definition 2.4.1. *Let N be the sample size and F_{ML} be the maximum likelihood fitting function. Then the χ^2 statistic is defined by*

$$\chi^2 = (N-1)F_{ML}. \tag{2.38}$$

The product in equation (2.38) follows the χ^2-distribution if the data is multivariate normal. Our null hypothesis equals: the sample covariance matrix and the model estimated matrix do not differ. Therefore instead of a significant χ^2, we want it to be non-significant. We want the test to not reject the null hypothesis.

However, there are some problems with the χ^2. The most important one is that it is highly sensitive to the sample size, since it is defined as $N-1$ multiplied with the minimized fitting function. Therefore it is biased to be significant with large sample sizes. For models with large sample sizes this can mean that the model is good, even though the χ^2 will indicate that the model does not fit the data.

2.4.1. Measures of absolute fit

Measures of absolute fit are model fit indices that are based on the assumption that a perfect model has no deviation between the observed and the model implied covariance matrices. They compare the model with the theoretically perfect model, the higher the value of the measure of absolute fit, the bigger the difference between the covariance matrices and thus the worse the fit.

The first measure of absolute fit we will discuss is the root mean square error of approximation.

Definition 2.4.2. *Let N be the sample size and df the degrees of freedom of the model. Then the root mean square error of approximation (RMSEA) is defined by*

$$RMSEA = \sqrt{\frac{\chi^2 - df}{df(N-1)}}.$$

Thus it measures the average lack of fit per degree of freedom. It ranges from 0 to 1, with smaller values indicating a better model fit. Literature suggests that a value of the RMSEA between 0 and 0.06 represents good model fit, a value between 0.06-0.08 a fair model fit and values above 0.08 poor fit.

The next measure of absolute fit we discuss is based on standardized residuals.

Definition 2.4.3. *Let p be the number of observed variables and let s and σ be the elements from the sample covariance matrix and the model estimated covariance matrix, respectively. The standardized root mean squared residual is defined by*

$$SRMR = \sqrt{\frac{\sum_j \sum_k \left(\frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} - \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} \right)^2}{p(p-1)/2}}.$$

In words, the SRMR is the average square root of the residual between the sample covariance matrix and the model estimated covariance matrix, which has been standardized to range between 0 and 1. By looking at definition (2.4.3) we indeed see that the squared residuals are taken in the numerator and are divided by $p(p+1)/2$ to take the average of those residuals. The acceptable range for the SRMR is between 0 and 0.08.

2.4.2. Measures of relative fit

Measures of relative fit compare the model to the worst possible model, the null-model. The null-model presumes that all variables are uncorrelated. So the measures of relative fit can be seen as goodness of fit measures, they tell us how much better the model is than the null-model. Therefore a higher value of the measure will indicate a better model fit.

The first measure of relative fit we will discuss is the comparative fit index.

Definition 2.4.4. *Let $d = \chi^2 - df$. Then the comparative fit index (CFI) is defined as*

$$CFI = \frac{d_{null} - d_{model}}{d_{null}}$$

The values range between 0 and 1, where values above 0.9 are considered to indicate a good fit. But since the CFI depends on the average size of the correlations in the data, the CFI will have a smaller value when the average correlation between the variables is not high.

The next measure of relative fit is very similar to the CFI and is often reported together with it.

Definition 2.4.5. *The Tucker-Lewis index (TLI) is defined by*

$$TLI = \frac{\left(\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{model}}{df_{model}}\right)}{\left(\frac{\chi^2_{null}}{df_{null}}\right)}.$$

Just like the CFI, values above 0.9 are considered to be a sign of a good model fit. The TLI will also not be high if the average correlation between the variables is not high, since it also depends on the average size of the correlations in the data.

2.5. Model modification

Unless the model immediately fits the data satisfactorily, the model has to be modified after evaluation. Therefore we are interested in the effects of making changes in model specification on the model fit. To that end, let us consider the following theorem.

Theorem 2.5.1. *Let f be a fitting function, let θ be the vector of free parameters and let $g = \frac{\partial f}{\partial \theta}$ denote the gradient vector of f. Let $E = \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta^2}\right]$ denote the matrix of expected second order derivatives of f. Then an estimate for the decrease in f when freeing a fixed parameter ($\hat{F}$) is given by:*

$$\hat{F} = \frac{\frac{1}{2}\hat{g_1^2}}{\mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] - \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^{\intercal} \hat{E}^{-1} \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]}. \tag{2.39}$$

Proof (see [4]). The Taylor expansion of f around $\hat{\theta}$ equals

$$f \approx \hat{f} + (\theta - \hat{\theta})^{\intercal}\hat{g} + \frac{1}{2}(\theta - \hat{\theta})^{\intercal}\hat{f}(\theta - \hat{\theta}), \tag{2.40}$$

where $\hat{f}$ and $\hat{g}$ are the values of f and g at the solution $\hat{\theta}$, respectively.
We want a previously fixed parameter to be set free, thus added to the vector of free paramaters. Let θ_1 be a new free parameter and let $\hat{\theta_1}$ be the previously fixed value of θ_1.
The Taylor expansion then equals

$$f \approx \hat{f} + \begin{bmatrix} \theta - \hat{\theta} \\ \theta_1 - \hat{\theta_1} \end{bmatrix}^{\intercal} \begin{bmatrix} \hat{g} \\ \hat{g_1} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \theta - \hat{\theta} \\ \theta_1 - \hat{\theta_1} \end{bmatrix}^{\intercal} \begin{bmatrix} \hat{E} & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right] \\ \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^{\intercal} & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] \end{bmatrix} \begin{bmatrix} \theta - \hat{\theta} \\ \theta_1 - \hat{\theta_1} \end{bmatrix}, \tag{2.41}$$

where $\hat{E}$ is the value of E at $\hat{\theta}$. We want to compute how much f in equation (2.41) will decrease. In order to show that, we find new estimates by minimizing f with respect to (θ, θ_1). Hence the equation

$$\begin{bmatrix} \frac{\partial f}{\partial \theta} \\ \frac{\partial f}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.42}$$

must be satisfied. Since $\hat{g} = 0$, equation (2.42) becomes

$$\begin{bmatrix} \frac{\partial f}{\partial \theta} \\ \frac{\partial f}{\partial \theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{g_1} \end{bmatrix} + \begin{bmatrix} \hat{E} & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right] \\ \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^\intercal & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] \end{bmatrix} \begin{bmatrix} \theta - \hat{\theta} \\ \theta_1 - \hat{\theta_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.43}$$

Solving (2.43) gives us

$$\begin{bmatrix} \theta - \hat{\theta} \\ \theta_1 - \hat{\theta_1} \end{bmatrix} = -\begin{bmatrix} 0 \\ \hat{g_1} \end{bmatrix} \begin{bmatrix} \hat{E} & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right] \\ \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^\intercal & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] \end{bmatrix}^{-1} \tag{2.44}$$

Finally, substituting equation (2.44) into (2.41) gives us the result

$$f_{min} = \hat{f} - \begin{bmatrix} 0 \\ \hat{g_1} \end{bmatrix}^\intercal \begin{bmatrix} \hat{E} & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right] \\ \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^\intercal & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \hat{g_1} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 \\ \hat{g_1} \end{bmatrix}^\intercal \begin{bmatrix} \hat{E} & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right] \\ \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^\intercal & \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \hat{g_1} \end{bmatrix} \tag{2.45}$$

$$= \hat{f} - \frac{\frac{1}{2}\hat{g_1^2}}{\mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] - \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^\intercal \hat{E}^{-1} \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]}. \tag{2.46}$$

Then $\hat{F}$ is given by $\hat{f} - f_{min}$ and hence equals

$$\hat{F} = \frac{\frac{1}{2}\hat{g_1^2}}{\mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_1^2}\right] - \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]^\intercal \hat{E}^{-1} \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta \partial \theta_1}\right]}. \tag{2.47}$$

□

A direct result of theorem (2.5.1) is the following corollary.

Corollary 2.5.1.1. *$\hat{F}$ is an estimate of how much the difference between the sample covariance matrix and the model estimated matrix will be reduced, and thus of how much the model fit will improve. $\hat{F}$ is also called a modification index.*

In practice we can let R (or other SEM programs) give us a list with all the modification indices, with suggestions of changes in model structure. Before making changes in the model structure, that are suggested by the modification indices, we should always check whether those changes make sense theoretically.

3

PRIME research

In this chapter a practical example of structural equation modeling will be given, in order to illustrate the use and the usefulness of SEM. We will use R to apply SEM on a hypothesized model with collected data. The results are not as good as we would like, but this chapter will still serve as a practical example of the theory discussed in Chapter 2, and as a demonstration of how to interpret results of SEM.

3.1. Research Plan

The PRogramme of Innovation in Mathematics Education [11] (PRIME) is redesigning mathematics courses for engineers. The aim of PRIME is a permanent programme at TU Delft intended to

- improve study results,
- improve connection between mathematics and engineering,
- increase students active participation and motivation.

We want to research the impact of the innovation on the quality of education and student success and motivation. The quality of education and student success and motivation are all not directly measurable, and we are interested in the relationships between them. Estimating those relationships would give us the insight we want. So we want to estimate relationships between unmeasured, thus latent, variables and therefore we will use structural equation modeling. To that end, we need to define the hypothesized model.

3.1.1. Research question

To research the impact of the innovation a research question was formulated:

What are the relationships between student perceptions of teacher cues in monitoring and scaffolding, satisfaction and frustration of needs for relatedness, autonomy and competence, motivation, and academic performance?

From this research question, the hypothesized model can be formulated. But first we will explain some of the psychological terms used in the model.

Relatedness The feeling of connection with others

Autonomy The ability to make choices according to one's own free will

Competence The ability to do something successfully

Autonomous motivation Motivation coming from oneself

Controlled motivation Motivation that is influenced by punishment or reward

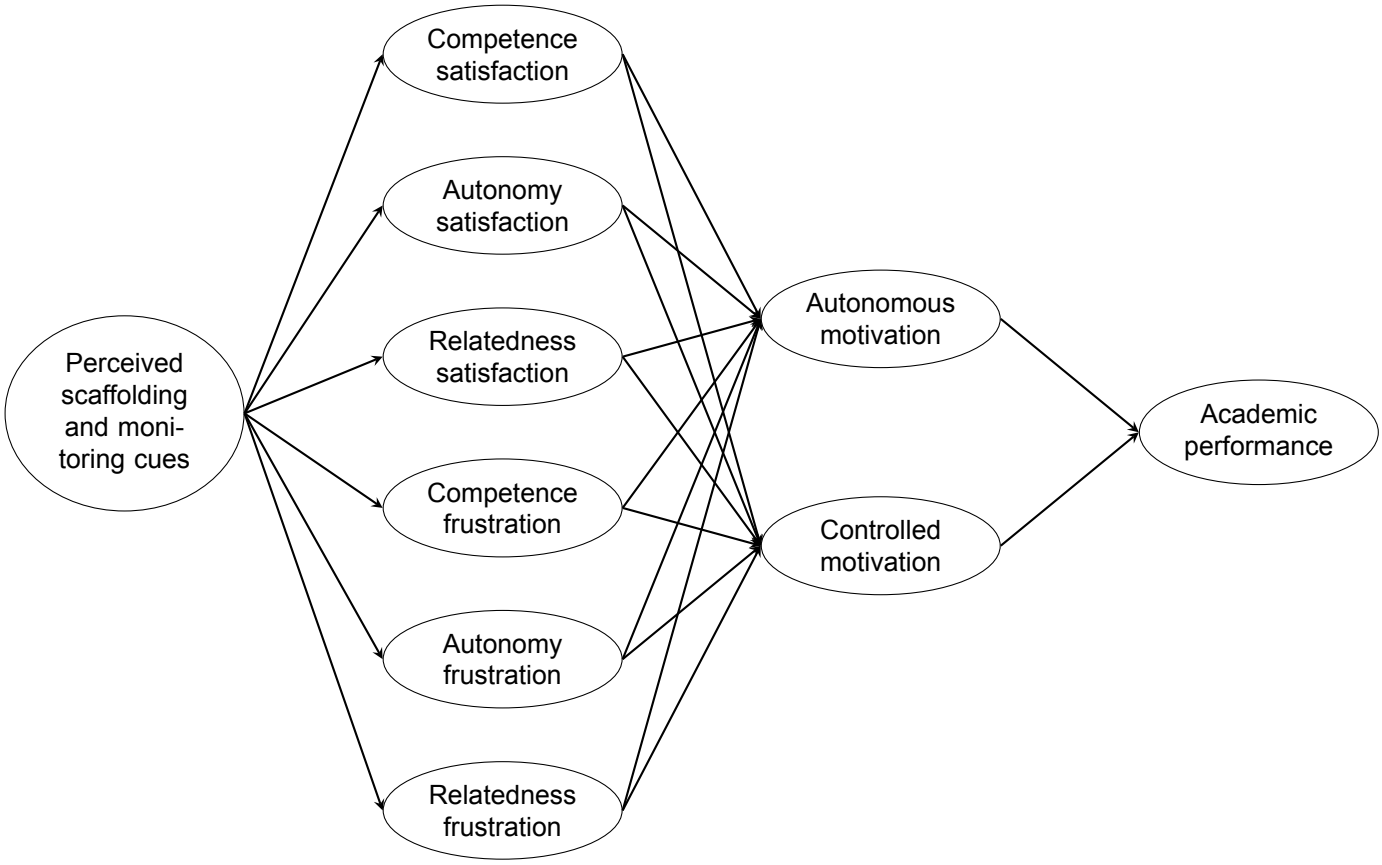


Figure 3.1: Path diagram corresponding to the research question

To clarify what the goal of the research question is, we will represent it as a path diagram. The latent variables in our model are: perceived scaffolding and monitoring cues, competence satisfaction, competence frustration, autonomy satisfaction , autonomy frustration, relatedness satisfaction, relatedness frustration, autonomous motivation, controlled motivation and academic performance.
The path diagram representation of the research question, which exists only of the latent variables, is shown in (3.1). In this path diagram we can see the hypothesized relations that we want to investigate. In the path diagram the observed variables are omitted, because of the size of the model.

3.1.2. Hypotheses

Several hypotheses were formulated for the research question and the associated path diagram. They describe what we expect of the relationships between the latent variables. The following hypotheses are of interest to us:

- Perceived scaffolding and monitoring cues are positively related to competence and relatedness satisfaction and negatively related to competence and relatedness frustration.
- Autonomy satisfaction is positively related to autonomous motivation and autonomy frustration is positively related to controlled motivation.
- Autonomy, competence and relatedness satisfaction are negatively related to controlled motivation and autonomy, competence and relatedness frustration are negatively related to autonomous motivation.
- Autonomous motivation is positively related to achievement.
- Controlled motivation is negatively related to achievement.

3.1.3. Data

Among students following a mathematics course, which was innovated by PRIME, a questionnaire was held. For every latent variable in (3.1), except for academic performance, several questions were composed and these questions serve as observed variables of those latent variables. The questionnaire consisted of 30 questions about teacher cues, 24 questions about relatedness, competence and autonomy, and 16 questions about motivation. Finally, to measure academic performance, past grades and the grades of the exams of the course were collected. In total, data of 220 students could be used for the research.

3.2. Formulating the model in R

Now that we know which hypothesized relationships we want to investigate, we can formulate the model in R. There are several R-packages which we can use to perform structural equation modeling. The R-package that we are going to use, a package called ”lavaan”, has its own syntax for describing the model. The model syntax is summarized in table (3.1).

Formula type	Operator	Meaning
latent variable definition	=∼	is measured by
regression	∼	is regressed on
(residual) (co)variance	∼∼	is correlated with

Table 3.1: Lavaan Model syntax

Using the model syntax and by looking at the path diagram (3.1), we can define the structural and the measurement part of the model in R. For the measurement part we use the questions from the questionnaire associated with the particular latent variable as observed variables, thus it is indicated for each latent variable by which observed variable it is being measured. The latent variable definitions using the model syntax can be seen below.

```
#latent variable definitions (measurement part)
SMC =~    S_M1 + S_M2 + S_M3 + S_M4 + S_M5
          + S_M11 + S_M12 + S_M13 + S_M14 + S_M15
          + S_M16 + S_M17 + S_M18 + S_M19 + S_M20
          + S_M21 + S_M22 + S_M23 + S_M24 + S_M25
          + S_M26 + S_M27 + S_M28 + S_M29 + S_M30

autonomous_motivation =~  Motivation2 + Motivation4 + Motivation7 + Motivation8
                          + Motivation11 + Motivation13 + Motivation15 + Motivation16
controlled_motivation =~  Motivation1 + Motivation3 + Motivation5 + Motivation6
                          + Motivation9 + Motivation10 + Motivation12 + Motivation14

relatedness_satisfaction =~ Needs3 + Needs9  + Needs15 + Needs21
relatedness_frustration  =~ Needs4 + Needs10 + Needs16 + Needs22

competence_satisfaction =~ Needs1 + Needs7 + Needs13 + Needs19
competence_frustration  =~ Needs2 + Needs8 + Needs14 + Needs20

autonomous_satisfaction =~ Needs5 + Needs11 + Needs17 + Needs23
autonomous_frustration  =~ Needs6 + Needs12 + Needs18 + Needs24

academic_performance =~  Grade_1 + Grade_2
                       + Final_grade + Past_performance_mathematics
```

Where SMC is the latent variable perceived scaffolding and monitoring cues.

In the structural part of the equation the relations between the latent variables are established. By looking at the path diagram (3.1) it is clear which latent variable is connected to which other latent variable. Hence the structural part in model syntax looks like this:

```
#regressions (structural part)
relatedness_satisfaction ~ SMC
relatedness_frustration  ~ SMC
competence_satisfaction  ~ SMC
competence_frustration   ~ SMC
autonomous_satisfaction  ~ SMC
autonomous_frustration   ~ SMC

autonomous_motivation ~ relatedness_satisfaction + relatedness_frustration
                        + competence_satisfaction + competence_frustration
                        + autonomous_satisfaction + autonomous_frustration
controlled_motivation ~ relatedness_satisfaction + relatedness_frustration
                        + competence_satisfaction + competence_frustration
                        + autonomous_satisfaction + autonomous_frustration

academicperformance ~ autonomous_motivation + controlled_motivation
```

To check whether the model is formulated in the way we want it, we can let R draw path diagram of the model and compare it to our hypothesized path diagram (3.1). Again we only show the latent variables because of the size of the model.

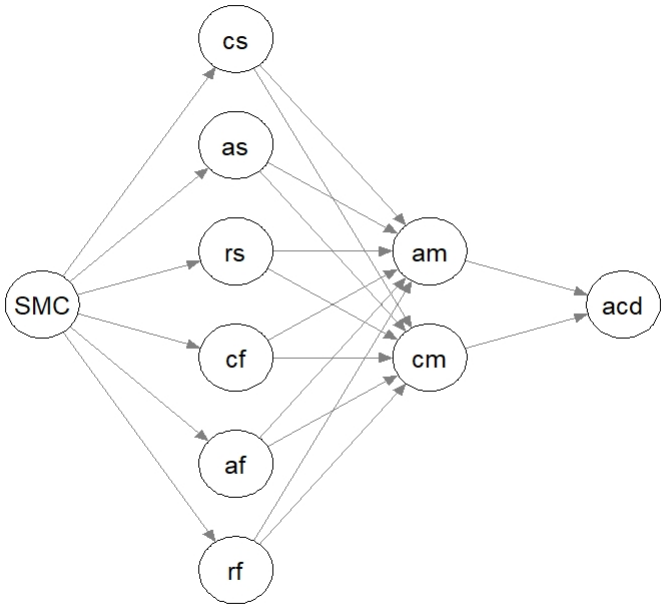


Figure 3.2: Path diagram of the model in R

We can see that the path diagram in (3.2) and the path diagram in (3.1) describe the same model and therefore the model is formulated in the way we want it.

3.3. Fitting the Model in R

Now that the model has been formulated, we are ready to fit the model in R. In this section we will discuss the process of identification, estimation, evaluation and modification of the model in R.

3.3.1. Model identification and estimation

The R-package has a function that estimates the model as described in section (2.3), it uses the maximum-likelihood fitting function and iteratively solves for the parameters. For the sake of identification, the function automatically fixes the factor loading of the first indicator of every latent variable to be equal to one. We should also check if the t-rule, as defined in (2.12) holds. Since we have a large number of observed variables, namely 74, it definitely holds:

$$t = 168 \leq \frac{1}{2}(74)(75) = 2775 \tag{3.1}$$

Now that we have confirmed that the model is identified we can let R fit the model, and plot the path diagram of the model with the calculated estimates of the links between the latent variables shown on the edges.

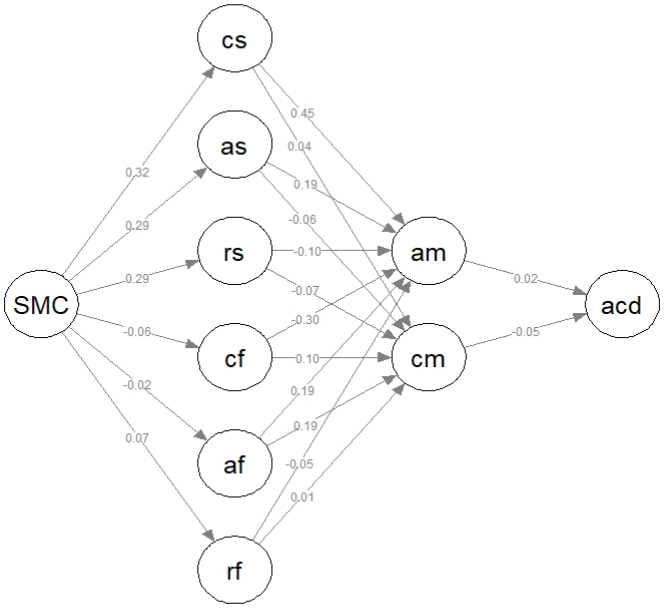


Figure 3.3: Path diagram of the model with standardized estimates

But before we can conclude anything from the estimates shown in (3.4) we have to see if the model fits the data well, otherwise the estimates might not be reliable.

3.3.2. Model evaluation and modification

After estimating the model we need to evaluate and modificate the model as described in sections (2.4) and (2.5). To that end we start with taking a look at the fit indices.

Fit index	Value
P-value (χ^2)	0.000
RMSEA	0.068
SRMR	0.110
CFI	0.714
TLI	0.704

Table 3.2: Values of the fit indices

In table (3.2) we can see the values of the fit indices that we are going to discuss. We start with the χ^2-statistic. The P-value indicates that the χ^2-statistic is significant, since clearly P $\leq$ 0.05. This means that the null hypothesis, the sample covariance matrix and the model estimated matrix do not differ, is rejected. Since we do not want the sample and model estimated covariance matrices to differ, the χ^2-statistic suggests that the data does not fit the model well. However as earlier discussed, for a large sample size the test is biased to be significant. Since we have a large sample size (220), the χ^2-statistic is biased. Hence we should not conclude anything from the χ^2-statistic only.
The root mean square error of approximation, as defined in (2.4.2), has a value of 0.068. This suggests a fair, but not as good as we want, model fit. However the values of the standardized root mean squared residual, the comparative fit index and the Tucker-Lewis index, as defined in (2.4.3)-(2.4.5), all indicate poor model fit, the SRMR being too high and the CFI and TLI being too low.
Since the model fit tests indicate that we need to improve the model fit, we are going to take a look at the modification indices. After checking if the suggested changes theoretically make sense, which in this case means we need to check if it makes sense that certain questions correlate, we can add the following residual covariances to the model formulation in R.

```
#residual covariances
 S_M19 ~~ S_M20
 S_M7 ~~ S_M8
 S_M3 ~~ S_M6
 S_M9 ~~ S_M10
 S_M3 ~~ S_M11
 S_M15 ~~ S_M16
 S_M11 ~~ S_M12
 S_M21 ~~ S_M24
 S_M24 ~~ S_M27
 S_M25 ~~ S_M27
 S_M24 ~~ S_M25
 S_M3 ~~ S_M14
 S_M3 ~~ S_M4
 S_M28 ~~ S_M30
 S_M9 ~~ S_M16
 S_M12 ~~ S_M18
 S_M14 ~~ S_M16
 S_M28 ~~ S_M29
 S_M2 ~~ S_M15
 S_M22 ~~ S_M23

 Motivation8 ~~ Motivation13
 Motivation13 ~~ Motivation16
 Motivation8 ~~ Motivation16
 Motivation6 ~~ Motivation10
 Motivation6 ~~  Motivation9
 Motivation5 ~~ Motivation12
 Motivation1 ~~ Motivation14
 Motivation7 ~~ Motivation15

 Needs9 ~~ Needs21
 Needs15 ~~ Needs21
 Needs3 ~~ Needs9
 Needs6 ~~ Needs18
```

After again fitting the model in R with the new model formulation, we can take a look at the new values of the fit indices.

Fit index	Value
P-value (χ^2)	0.000
RMSEA	0.057
SRMR	0.107
CFI	0.802
TLI	0.793

Table 3.3: Values of the fit indices

The first thing we notice from table (3.3) is that the model fit has improved, since the values of the RMSEA and SRMR are lower than before and the values of the CFI and TLI are higher than before. However, the values of the SRMR, CFI and TLI are still not acceptable and the χ^2-statistic is still significant. Since there are no suggestions of the modification indices left that theoretically make sense, we need to look elsewhere to improve the model fit.
By taking a look at the summary we notice something in the measurement part of the latent variable SMC (scaffolding and monitoring cues).

SMC =~	Estimate		
S_M1	1.000	S_M16	1.153
S_M2	0.859	S_M17	0.412
S_M3	0.952	S_M18	0.925
S_M4	0.920	S_M19	0.491
S_M5	0.415	S_M20	0.499
S_M6	1.090	S_M21	0.746
S_M7	1.349	S_M22	0.321
S_M8	1.067	S_M23	0.362
S_M9	1.549	S_M24	0.455
S_M10	1.259	S_M25	0.301
S_M11	0.929	S_M26	1.007
S_M12	0.859	S_M27	0.213
S_M13	0.490	S_M28	0.240
S_M14	1.068	S_M29	0.315
S_M15	1.011	S_M30	0.338

Table 3.4: Factor loadings of the latent variable SMC

In table (3.4) we can see that some of the factor loadings are low compared to the others, and this could be causing problems for the model. A certain factor loading being low can be interpreted as the corresponding observed variable not being a good indicator for the latent variable. Since the latent variable SMC has a high number of observed variables, we could try deleting some of the observed variables with low factor loadings without having problems with identification and theoretically the model would still make sense. Removing questions 5, 17, 22-25 and 27-30 from the measurement part and removing the corresponding residual variances, and fitting the new model gives us the following new values of the fit indices.

Fit index	Value
RMSEA	0.055
SRMR	0.109
CFI	0.841
TLI	0.832

Table 3.5: Values of the fit indices

We can see that the RMSEA, CFI and TLI have improved. However again the values of the fit

indices, except for the RMSEA, indicate that the model still does not fit the data well. This could mean that the model we want to test just does not fit the data, and thus that we should change our hypothesized model. But it could also mean that the model does fit the data, only the tests show otherwise. Since there is no immediate way to improve our model fit, we are going to stop here with the model evaluation and modification.

3.4. Results

With the collected data and the hypothesized model a better model fit, according to the fit indices, is not possible. But we can still investigate what the obtained estimates tell us and interpret them.
Below the standardized estimates of the links between the variables of the last fitted model of section (3.3.2) are shown on the edges of the path diagram and in tables.

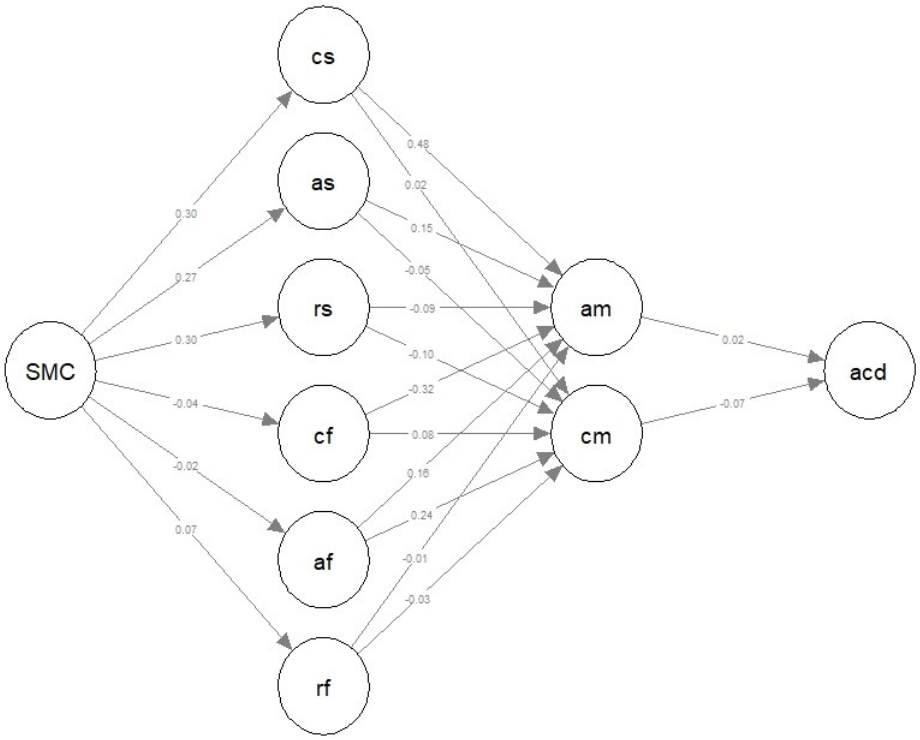


Figure 3.4: Path diagram of the model with standardized estimates

Table 3.6: Standardized estimates of the regressions

Table 3.7

Regression	Standardized estimate
rs ~ SMC	0.301
rf ~ SMC	0.071
cs ~ SMC	0.304
cf ~ SMC	-0.041
as ~ SMC	0.272
af ~ SMC	-0.018

Table 3.8

am ~	Standardized estimate
rs	-0.089
rf	-0.014
cs	0.481
cf	-0.317
as	0.148
af	0.161

Table 3.9

cm ~	Standardized estimate
rs	-0.096
rf	-0.032
cs	0.024
cf	0.080
as	-0.047
af	0.244

Table 3.10

academicperformance ~	Standardized estimate
am	0.016
cm	-0.073

The standardized estimates show how many standard deviations the variable will change, per standard deviation increase of the predictor variable. For example, a standard deviation increase of the latent variable scaffolding and monitoring cues corresponds with a 0.301 standard deviation increase

of the latent variable relatedness satisfaction and with a 0.041 standard deviation decrease of the latent variable competence frustration. Thus the higher the standardized estimate, the stronger the relationship between the corresponding variables.

3.4.1. Interpretation of the estimates

Now that we have the estimates of the relationships between the latent variables, we can try to answer the research question formulated in section (3.1.1). To that end, we will describe the relationships between the latent variables.
We start with the influence of perceived scaffolding and monitoring cues on competence, autonomy and relatedness satisfaction and frustration. In table (3.7) the standardized estimates of the influences of perceived scaffolding and monitoring are shown. The strongest positive relationships are with relatedness satisfaction, competence satisfaction and autonomy satisfaction. Furthermore, perceived scaffolding and monitoring has a less strong positive influence on relatedness frustration, and a negative influence on competence frustration and autonomy frustration.

Now we take a look at the relationships between competence, autonomy and relatedness satisfaction and frustration and autonomous motivation, shown in table (3.8). Relatedness satisfaction, relatedness frustration and competence frustration have a negative influence on autonomous motivation, where the relationship with competence frustration is clearly the strongest. Autonomous motivation has positive influences from autonomy satisfaction, autonomy frustration and competence satisfaction, of which the latter has the strongest influence.

Then from table (3.9), the influence of competence, autonomy and relatedness satisfaction and frustration on controlled motivation. Relatedness satisfaction, relatedness frustration and autonomy satisfaction have a negative influence on controlled motivation. Competence satisfaction, competence frustration and autonomy frustration have a positive influence on controlled motivation, where autonomy frustration has the strongest influence.

Finally, we take a look at the influence of autonomous motivation and controlled motivation on academic performance, of which the standardized estimates are shown in table (3.10). We see that autonomous motivation has a positive influence on academic performance and controlled motivation has a negative influence on academic performance, where the relationship with controlled motivation is stronger.

With the established relationships between the variables, we can determine whether the results are as expected by comparing them with the hypotheses defined in section (3.1.2).

We start with the first hypothesis:

- Perceived scaffolding and monitoring cues are positively related to competence and relatedness satisfaction and negatively related to competence and relatedness frustration.

Perceived scaffolding and monitoring cues has a positive influence on relatedness frustration, which is not as expected. However, the other relationships described in the hypothesis above correspond with the results, and thus are as expected.

Now the second hypothesis:

- Autonomy satisfaction is positively related to autonomous motivation and autonomy frustration is positively related to controlled motivation.

Both of the relations described in this hypothesis correspond with the results.

Then the hypothesis:

- Autonomy, competence and relatedness satisfaction are negatively related to controlled motivation and autonomy, competence and relatedness frustration are negatively related to autonomous motivation.

Here two of the statements in the hypothesis do not correspond with the results, because competence satisfaction is positively related to controlled motivation and autonomy frustration is positively related

to autonomous motivation. The other statements do correspond with the results, thus are as expected.

Finally, the last two hypotheses:

- Autonomous motivation is positively related to achievement.
- Controlled motivation is negatively related to achievement.

Both of the statements are true, thus the relationships between motivation and academic performance are as expected.

Even though the model fit tests indicate that parameter estimates might not be reliable, the greater part of the results correspond with what we expected.

4

Conclusion

In this thesis we have attempted to determine the mathematical foundation of structural equation modeling, a method to estimate links between latent variables, and to illustrate the use of this method.

We started with studying the mathematical theory of every step involved in the procedure of SEM, and where it was needed we introduced mathematical formulation of the theory. We saw that there are multiple ways to describe a hypothesized model, with formulation as equations the most mathematically correct one. Then we looked at certain identification conditions that must be met in order to estimate the parameters, they ensure whether a solution can be determined. We have seen that the parameters can be estimated by minimizing the difference between two covariance matrices, with the help of a fitting function. After estimation, there has to be checked if the data indeed shows what has been hypothesized, and to that end several tests to verify this were introduced. Finally, in the case the model does not represent the data, we showed a method to try to solve this problem.
The five above mentioned steps, discussed in Chapter 2, form the mathematical formulation of how to estimate links between latent variables using structural equation modeling.

To illustrate the use of structural equation modeling, we applied the method to collected data. We followed the steps we studied beforehand, and estimated the links between the latent variables of the hypothesized model. However, the hypothesized model did not represent the collected data, according to model fit tests. From this finding, we can see that determining the hypothesized model is of great importance when applying structural equation modeling. Finally, we showed how to interpret the estimated links between the latent variables.

5 Discussion

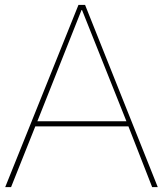
In this chapter some of the complications we have run into during this research will be discussed and in addition we will give suggestions for future research.

5.1. Fit indices

In section (2.4) we talked about evaluating the model to see if the model fits the data, and to test this we introduced some measures of fit. However, there exists controversy about measures of fit.
First of all, controversy about the cutoff values of the fit indices. There are a lot of different opinions over when values of the fit indices suggest a certain type of model fit, good, fair or bad. In literature different cutoff values or 'rules of thumb' are suggested to assess the model fit. Hence it is not completely clear which values to use.
Second of all, controversy about different fit indices suggesting different types of model fit. There are a lot of fit measures available, and often they can differ in the assessment of the model fit. It is unclear what conclusions you can make when fit measures offer different evaluations of model fit. A problem that comes with this is that one could compute many different fit indices, and only report the ones that suggest the result that they want.
Lastly, controversy about which fit indices to use. Since there are a lot of fit measures, one has to pick which fit indices to look at. But there is disagreement on which fit indices should be considered for model evaluation.
The above mentioned problems with fit indices give a good reason to further study what cutoff values should be used, how we can interpret tests giving different evaluations of model fit, and what fit measures should be used to evaluate the model.

5.2. Result PRIME research

From section (3.3.2) we concluded that the hypothesized model used for the PRIME research did not represent the collected data. Thus the estimates of the links between the variables and the interpretation of them, that we gave in section (3.4), might not be reliable. We saw no other solution to this problem, than to change the hypothesized model. But since we did not have the psychological knowledge to obtain a new psychologically correct hypothesized model, our analysis stopped there. But even if we did obtain a new hypothesized model, we could have run into the same problem. Therefore a suggestion for future research is to further study if there are certain conditions a hypothesized model should meet in order to have a greater chance at a good model fit. For example, less complex hypothesized models might be more likely to represent the data. Then for the PRIME research we could use a hypothesized model with less variables, or we could test several smaller hypothesized models.



R-code

```
library(lavaan)
library(semPlot)
library(readxl)

my_data <- read_excel("Data_final.xlsx")

researchmodel<-'
#latent variable definitions (measurement part)
SMC =~     S_M1 + S_M2 + S_M3 + S_M4
           + S_M11 + S_M12 + S_M13 + S_M14 + S_M15
           + S_M16 + S_M18 + S_M19 + S_M20
           + S_M21 + S_M26

autonomous_motivation =~  Motivation2 + Motivation4 + Motivation7 + Motivation8
                          + Motivation11 + Motivation13 + Motivation15 + Motivation16
controlled_motivation =~  Motivation1 + Motivation3 + Motivation5 + Motivation6
                          + Motivation9 + Motivation10 + Motivation12 + Motivation14

relatedness_satisfaction =~ Needs3 + Needs9  + Needs15 + Needs21
relatedness_frustration  =~ Needs4 + Needs10 + Needs16 + Needs22

competence_satisfaction =~ Needs1 + Needs7 + Needs13 + Needs19
competence_frustration  =~ Needs2 + Needs8 + Needs14 + Needs20

autonomous_satisfaction =~ Needs5 + Needs11 + Needs17 + Needs23
autonomous_frustration  =~ Needs6 + Needs12 + Needs18 + Needs24

academic_performance =~  Grade_1 + Grade_2
                        + Final_grade + Past_performance_mathematics
#define structural part
#regressions (structural part)
relatedness_satisfaction ~ SMC
relatedness_frustration  ~ SMC
competence_satisfaction  ~ SMC
competence_frustration   ~ SMC
autonomous_satisfaction  ~ SMC
autonomous_frustration   ~ SMC

autonomous_motivation ~ relatedness_satisfaction + relatedness_frustration
                        + competence_satisfaction + competence_frustration
                        + autonomous_satisfaction + autonomous_frustration
controlled_motivation ~ relatedness_satisfaction + relatedness_frustration
                        + competence_satisfaction + competence_frustration
```

```
                        + autonomous_satisfaction + autonomous_frustration

academicperformance ~ autonomous_motivation + controlled_motivation

#residual covariances
 S_M19 ~~ S_M20
 S_M7 ~~ S_M8
 S_M3 ~~ S_M6
 S_M9 ~~ S_M10
 S_M3 ~~ S_M11
 S_M15 ~~ S_M16
 S_M11 ~~ S_M12

 Motivation8 ~~ Motivation13
 Motivation13 ~~ Motivation16
 Motivation8 ~~ Motivation16
 Motivation6 ~~ Motivation10
 Motivation6 ~~  Motivation9
 Motivation5 ~~ Motivation12
 Motivation1 ~~ Motivation14
 Motivation7 ~~ Motivation15

 Needs9 ~~ Needs21
 Needs15 ~~ Needs21
 Needs3 ~~ Needs9
 Needs6 ~~ Needs18
'


fit<-sem(researchmodel, as.data.frame(my_data))
summary(fit, standardized = TRUE, fit.measures=TRUE)
semPaths(fit, structural= TRUE, reorder = F, whatLabels = "std",  latents=c("SMC",
"rf","af","cf","rs","as","cs","cm","am",
"academicperformance"), rotation = 2 ,  layout = "tree2",
residuals = F, edge.label.position = c(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.3,
0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.5,0.5))
modificationindices(fit, sort. = T)
```

Bibliography

[1] Madansky, A. *Instrumental variables in factor analysis*. Psychometrika 29, 105–113 (1964).

[2] Hägglund, G. *Factor analysis by instrumental variables methods*. Psychometrika 47, 209–222 (1982).

[3] Jöreskog, K.G. *Structural analysis of covariance and correlation matrices*. Psychometrika 43, 443–477 (1978).

[4] Sörbom, D. *Model modification*. Psychometrika 54, 371–384 (1989).

[5] Wang, Jichuan, and Xiaoqian Wang. *Structural equation modeling: Applications using Mplus*. John Wiley & Sons, (2012).

[6] Kline, Rex B. *Principles and practice of structural equation modeling*. Guilford publications, (2015).

[7] Hoyle, Rick H., ed. *Handbook of structural equation modeling*. Guilford press, (2012).

[8] Bollen, K. A. *Structural equations with latent variables*. New York: Wiley, (1989)

[9] Bollen, Kenneth A., and J. Scott Long. *Testing structural equation models*. Vol. 154. Sage, (1993).

[10] Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*. Third edition. John Wiley & Sons, (2003).

[11] PRIME, TU Delft. URL https://www.tudelft.nl/eemcs/the-faculty/departments/applied-mathematics/education/prime/