Estimation of a Multilevel Item Response Theory Model with a Latent Interaction Effect using an EM Algorithm

Dissertation
zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von
Tim Fabian Schaffland
aus Köln

Tübingen
2021

1. Betreuer:	Prof. Dr. Augustin Kelava
2. Betreuer:	Prof. Dr. Martin Biewen

Tag der mündlichen Prüfung:	06.07.2021
Dekan:	Prof. Dr. Josef Schmid
1. Gutachter:	Prof. Dr. Augustin Kelava
2. Gutachter:	Prof. Dr. Martin Biewen

Contents

Danksagung

Nach intensiver Arbeit an meiner Dissertation ist sie nun abgeschlossen und es ist endlich an der Zeit meinen Dank an alle auszusprechen, die mich in den letzten Jahren unterstützt haben.

Zuallererst möchte ich mich herzlich bei meinem Betreuer Professor Augustin Kelava für die gemeinsame Auswahl des Promotionsthemas bedanken. Ich konnte mich jederzeit mit Fragen und Problemen an ihn wenden, auch wenn es mal nicht um die Dissertation ging. Besonders unsere fachlichen Diskussionen haben mir immer weitergeholfen. Er hat mir viele Erfahrungen auf Konferenzen und einem Forschungsaufenthalt ermöglicht und für all das bin ich ihm sehr dankbar.

Außerdem, danke ich meinem Zweitgutachter Professor Martin Biewen für seine Unterstützung.

Ohne meinen Freund und Kollegen Stefano wäre diese Arbeit möglicherweise nie fertig geworden. Egal zu welcher Uhrzeit (doch wenn möglich nach 10 Uhr morgens) und egal an welchem Wochentag, er hat mir immer mit gutem Rat zur Seite gestanden. Nicht nur für seine zahlreichen Korrekturvorschläge an der Arbeit bin ich ihm überaus dankbar.

Besonderer Dank gilt meiner Frau Marlene, die mich in all der Zeit ertragen hat und mich trotz allem sogar mittendrin geheiratet hat.

Ich möchte mich aber auch bei allen meinen Freundinnen und Freunden bedanken. Bei Laura für die mentale Unterstützung in schweren Zeiten und für das gemeinsame Fußballschauen der 4. bulgarischen Liga. Bei Nele und Heide, die mir den Start in Tübingen erleichtert haben. Bei Johanna für die schönen Abendessen und Spaziergänge um den Kopf frei zu bekommen. Bei Julia dafür, dass sie eine wunderbare erste Büropartnerin war, mit der man auch – nach Feierabend natürlich – im Büro Skat spielen kann. Genauso möchte ich mich bei Eike für die fortgeführten Doppelkopfabende zu Hause bedanken. Bei Martin für die Anekdoten, die mich immer noch überraschen können. Bei Pascal für die leider viel zu wenigen Kneipenabende. Bei Amelie für die Mitgründung unseres Fanclubs. Bei Cora für unvergessene Festival-Erinnerungen. Bei Jonas, der mich immer wieder motivert hat. Bei Sebi für die immerwährende Unterstützung in langen Telefonaten. Bei Andy für den Austausch von Erfahrungen während der Promotion.

Zu guter Letzt danke ich meiner Familie, im Besonderen meinen Eltern und meiner Schwester, die immer für mich da sind.

Danke an Euch alle für Eure Unterstützung!

List of Abbreviations

BFGS	Broyden–Fletcher–Goldfarb–Shanno
EM	Expectation-Maximization algorithm
FIML	Full Information Maximum Likelihood
GHQ	Gauss-Hermite quadrature
GLLAMM	Generalized Linear Latent And Mixed Model
GLLVM	General Linear Latent Variable Model
GLM	Generalized Linear Model
GPCM	Generalized Partial Credit Model
GRM	Graded Response Model
IRS	Item Response Surface
IRT	Item Response Theory
JMLE	Joint Maximum Likelihood Estimation
LMS	Latent Moderated Structural Equations
MCMC	Markov Chain Monte Carlo approach
MHRM	Metropolis-Hastings-Robinson-Monroe algorithm
MINoLEM	Multilevel IRT with Nonlinear Latent variable Effects Model
MLE	Maximum Likelihood Estimation
ML-SEM	Multilevel Structural Equation Model
MML	Marginal Maximum Likelihood
OECD	Organisation for Economic Co-operation and Development
PCM	Partial Credit Model
PISA	Programme for International Student Assessment
RMSE	Root Mean Squared Error
SEM	Structural Equation Model
WLE	Weighted Likelihood Estimation
WLS	Weighted Least Square
WLSMV	Weighted Least Square Mean and Variance adjusted

Abstract

The size and therefore the complexity of collected datasets has been growing over time as computational capacities increase. Therefore, estimation methods that can take this complexity into account are needed. One such complex dataset is the Programme for International Student Assessment (PISA) by the Organisation for Economic Co-operation and Development (OECD), which is carried out to measure reading, mathematics, and science knowledge of 15-year-old students. PISA is conducted in different countries with different educational systems. Countries are ranked according to their students' performance, which can have direct political consequences for the educational system, especially in countries with lower ranks than their self-image would dictate.

Latent abilities are estimated in the PISA test using a 2PL model from the Item Response Theory (IRT) family. Before 2015, however, the Rasch model was used to describe the data until studies could show that the rankings change if more complex (and more plausible) models are used to analyze the PISA datasets. Kreiner and Christensen (2014), for example, demonstrated the usefulness of the inclusion of differential item functioning.

In this thesis, a multilevel IRT model with nonlinear latent variable effects model (MINoLEM) is presented. An estimation procedure based on the Expectation-Maximization algorithm is deduced. The accuracy of this estimation approach will be proven in several simulation studies and its usefulness will be shown through comparisons to other IRT software with the potential to include multilevel structures or nonlinear latent variable effects. The applicability of the MINoLEM estimation technique to real data is demonstrated by re-examining a PISA dataset and showing that latent interaction effects can be found.

Chapter 1

Introduction

Today, latent variables are very well known and widely used in several fields. However, the first mention of the concept – to the best of our knowledge today – came from Francis Galton (1888)[1]. According to Bollen (2002), the concept is "at least as old as religion". Nevertheless, Spearman (1904) was the first to formalize the idea of underlying (hidden) factors that explain the correlation between two variables:

> *"Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. [...] co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body."*

Spearman realized that people should score similarly on two different tests of mental abilities, thus suggesting the notion of the g-factor. However, since the two test results were often not identical, he assumed they could be explained as the sum of a "common factor" g and an individual component,

[1] This was brought to the attention of Bartholomew, Knott, and Moustaki (2011) by John Aldrich of the University of Southampton.

which he called the "specific factor". In the 1930s, Thurstone (1931, 1935, 1947) picked up Spearman's and also Binet's work and generalized it for the psychology context to include more than one common factor (van der Linden, 2016a). Since then, latent variable modeling has been further developed and refined in many different fields – including psychology, the social sciences, economics, medicine, and machine learning/artificial intelligence – with a great variety of different applications. As an unfortunate consequence, different names are used in different contexts. Typical examples are unobservable variable, abstract concept, random effect and still (common) factor (Bollen, 2002).

Latent variable models are widely used and probably familiar to most readers of this thesis. Consequently, this chapter will be used to introduce the main concepts that are relevant to this thesis and provide the necessary definitions. The readers will be referred to other publications for more detailed introductions.

In this first chapter, latent variables are formally introduced in Structural Equation Modeling (SEM). IRT – the main approach applied in this thesis – is presented by describing several important models. Multilevel models are then explained within the SEM and IRT contexts. Subsequently, Generalized Linear Latent And Mixed Models (GLLAMM) are introduced, which represent a (partly) unifying framework for multilevel models, IRT, and SEM. Throughout, the connection between IRT, SEM, and GLLAMM is explored and IRT is framed from different perspectives. Then, the Expectation-Maximization algorithm (EM) is explained, which is a powerful tool for estimating latent variable models. It enables the statistician to interpret model parameters as missing data and it helps avoid the need to compute the immensely complex observed likelihood function.

In the second chapter, a multilevel IRT with Nonlinear Latent variable Effects Model (MINoLEM) will be presented and an EM algorithm will be deduced to estimate this model. The implementation of the approach will be discussed as well as numerical aspects that need to be dealt with.

In the third chapter, the convergence of the model estimates to the true values will be shown in several simulation studies. Subsequently, the approach will be applied to a PISA dataset to show its applicability and usefulness in real-world data analyses.

Finally, the results of the thesis will be discussed. Limitations of the given approach will be examined as well as aspects of further research.

1.1 Structural Equation Models

According to Bartholomew et al. (2011), there are two main reasons for introducing latent variables in a model. First, “there may be variables of interest that either were not measured, or cannot be measured without error” (Spirtes, 2001). Examples are “mathematical ability” or “business confidence”, which are abstract constructs that cannot be measured directly. Nevertheless, statisticians would like to treat them as if they were measurable quantities. This requires a mapping of numbers to these abstract constructs. In practice, a construct is first clearly defined theoretically and then operationalized. That means that measurable indicators are chosen which (only) depend on the latent variable and follow the given theoretical definition of the latent variable.

The second reason for introducing latent variables is “that the latent variable model may be a more parsimonious representation of the distribution over the observed variables than any model without latent variables” (Spirtes, 2001). In other words, latent variables can help to reduce the dimensionality of a model in the sense that “the information contained in the interrelationships of many variables can be conveyed, to a good approximation, in a much smaller set” (Bartholomew et al., 2011), which is the broad idea behind factor analysis or data reduction techniques in general.

These two reasons essentially represent informal definitions of a latent variable as either an unmeasurable or unobservable variable or as a means of reducing data. Bollen (2002) mentions a third informal definition as a hypothetical construct, in that latent variables are not real but only exist as a thought experiment.

In his search for a more inclusive definition of a latent variable, Bollen (2002) also discusses two mathematical concepts that involve latent variables. First, local independence assumes that the connection between measured variables is the result of one or more latent variables that influence those observed variables. Local independence can formally be written as

$$P(X_1, \ldots, X_m) = P(X_1|\boldsymbol{\xi}) \cdots P(X_m|\boldsymbol{\xi})$$

where $X_1, \ldots, X_m$ are the observed variables and $\boldsymbol{\xi}$ is a vector of latent variables. Local independence provides insight into the usage of latent variables and describes them as a common source of variance for manifest variables.

Secondly, a latent variable can be seen as the expected value of the observations

$$T_i = \mathbb{E}\left[X_i\right].$$

This approach is adopted in Classical Test Theory, where T_i is not called a latent variable but the true score that would result if infinitely many observations were to be examined.

Bollen (2002) proposes that “a latent random (or nonrandom) variable is a random (or nonrandom) variable for which there is no sample realization for at least some observations in a given sample.” While he states that this definition is not necessarily new, it formalizes the idea that a random variable has no specific value – at least in a certain context. It also opens up the possibility of a broader interpretation of latent variables, like seeing latent variables as missing values. This point of view will be adopted later when exploring the EM algorithm. A more detailed discussion of these definitions is given in Bollen (2002).

All of these definitions have their advantages and disadvantages, but most importantly, the latent variable has to be clearly defined and operationalized. SEM is one example technique for linking measurable indicators to latent variables. This framework includes both reasons given above for the introduction of latent variables in the form of confirmatory factor analysis and exploratory factor analysis.

In the following paragraphs, the basic SEM approach is introduced and how nonlinear latent variable models are handled in this context is briefly discussed. For a more detailed introduction, the reader is referred to Bollen (1989).

1.1.1 The Basic Model

SEM is a combination of two aspects – the measurement model and the structural model. The measurement model is the mathematical approach to operationalizing the latent variables described above. Here, the collection of m observable indicators (mathematically random variables), also called manifest variables, will be denoted by $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)'$. The latent variable will be written as $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)'$. Since the latent variables cannot be observed directly, they are measured with some error using the measurement model:

$$\boldsymbol{X} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{\tau}$ is a $(m \times 1)$ vector of intercepts for $\boldsymbol{X}$, $\boldsymbol{\Lambda}$ is a $(m \times n)$ factor loading matrix giving the impact of $\boldsymbol{\xi}$ on $\boldsymbol{X}$, and $\boldsymbol{\epsilon}$ is a $(m \times 1)$ vector of measurement errors. It is usually assumed that the errors are independent and identically distributed and follow a normal distribution with mean zero. This measurement model describes the operationalization of all n latent variables in $\boldsymbol{\xi}$ and allows for so-called cross-loadings, which are loadings for manifest variables that are influenced by more than one latent variable.

The structural model, which stems from path analysis, captures the relations between several latent variables. In SEM, a distinction is made between exogenous variables (not influenced by other latent variables) and endogenous variables (influenced by other latent variables), which are denoted by $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{n_1})'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_{n_2})'$, respectively, where $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are vectors depicting n_1- and n_2-dimensional latent constructs:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

where $\boldsymbol{\alpha}$ is an n_2-dimensional intercept vector, $\boldsymbol{\Gamma}$ is the (n_2 x n_1) coefficient matrix giving $\boldsymbol{\xi}$'s impact on $\boldsymbol{\eta}$, and $\boldsymbol{\zeta}$ is the n_2-dimensional disturbance variable for $\boldsymbol{\eta}$ with $\mathbb{E}(\boldsymbol{\zeta}) = 0$ and $Cov(\boldsymbol{\zeta}, \boldsymbol{\xi}') = 0$. The two latent variables each have measurement models as in (1.1).

Figure 1.1 shows a simple path diagram with one exogenous latent variable ξ, one endogenous latent variable η with three manifest variables each. The indicators of η are labeled Y_1, Y_2, and Y_3 with measurement errors δ_1, δ_2 and δ_3 to simplify the distinction between the respective measurement models.

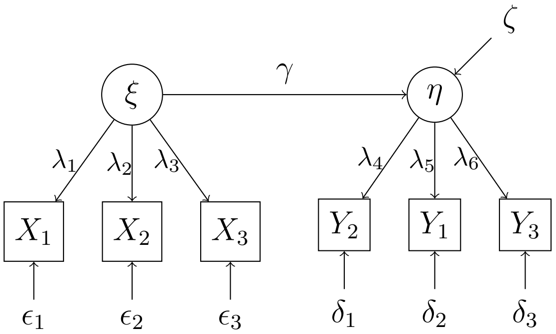

Figure 1.1: Path diagram of a simple Structural Equation Model with two latent variables ξ and η with three manifest variables each. The latent variable η is linearly dependent on ξ.

Model parameters can be estimated by considering the difference between the model implied variance-covariance matrix $\boldsymbol{\Sigma}$ and the observed variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ (Bollen, 1989).

1.1.2 Influence of Nonlinear Latent Variables

Having introduced the common notation, nonlinear latent variables will be discussed in this section. Over the years, SEM has been extended in many directions. For the purpose of this thesis, it is interesting to see how SEM can incorporate latent variables with a nonlinear influence.

Approaches to Include Nonlinearity

Within the parametric framework, several approaches have been developed, including product-indicator approaches (e.g., Kenny and Judd (1984), Jöreskog and Yang (1996), Bollen (1995, 1996), Ping (1995), Marsh, Wen, and Hau (2004), Kelava and Brandt (2009)), distribution-analytic approaches (e.g., Klein and Moosbrugger (2000), Klein and Muthen (2007)), moment-based approaches (e.g., Mooijaart and Bentler (2010)), and Bayesian approaches (e.g., Arminger and Muthen (1998), Lee (2007), Song and Lee (2007)).

One class of models that have been developed are semiparametric structural equation models, which use mixtures of linear structural equation models to approximate the nonlinearity in the data (e.g., Arminger and Stein (1997), Arminger, Stein, and Wittenberg (1999), Bauer (2005), Jedidi, Jagpal, and DeSarbo (1997), Kelava and Brandt (2014), and Kelava, Nagengast, and Brandt (2014)). The advantage of these models is their applicability without having to specify the function with which the latent variable nonlinearly influences the model. However, this is also a disadvantage, since it is not possible to quantify the nonlinear influence of the latent variable within this framework.

The nonparametric approach by Kelava, Kohler, Krzyzak, and Schaffland (2017) introduces a model in which the distribution of the latent variables is not restricted and potentially nonlinear connections between the latent variables can be estimated using nonlinear or nonparametric regression methods.

Here, the Latent Moderated Structural Equations (LMS) approach by Klein and Moosbrugger (2000) will be presented in more detail since it inspired the estimation approach that will be introduced later for a multilevel IRT model with nonlinear latent variable effects.

Latent Moderated Structural Equations approach

Klein and Moosbrugger (2000) assume a structural model with possible quadratic and interaction effects of the exogenous variable $\boldsymbol{\xi}$ on the endogenous variable η

$$\eta = \alpha + \boldsymbol{\Gamma\xi} + \boldsymbol{\xi}'\boldsymbol{\Omega\xi} + \zeta \tag{1.2}$$

where η is a 1-dimensional latent endogenous variable, α is an intercept term, $\boldsymbol{\xi}$ is a $(n \times 1)$ vector of latent exogenous variables, $\boldsymbol{\Gamma}$ is the $(1 \times n)$ coefficient vector giving $\boldsymbol{\xi}$'s impact on η, $\boldsymbol{\Omega}$ is the $(n \times n)$ coefficient matrix giving the impact of the product terms $\xi_i\xi_j$ $(i < j)$ on η (which is assumed to be an upper triangular matrix), and ζ is the disturbance variable. The measurement models of the latent variables are defined as in equation (1.1) with notation as in Figure 1.1. Under the usual assumptions that

- $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$,
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ and $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\delta)$ with $\boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{\Sigma}_\delta$ diagonal,
- $Cov(\boldsymbol{\epsilon}, \boldsymbol{\delta}) = Cov(\boldsymbol{\epsilon}, \boldsymbol{\xi}) = Cov(\boldsymbol{\epsilon}, \eta) = Cov(\boldsymbol{\delta}, \boldsymbol{\xi}) = Cov(\boldsymbol{\delta}, \eta) = 0$, and
- $\zeta \sim \mathcal{N}(0, \sigma_\zeta)$ with $Cov(\zeta, \boldsymbol{\xi}) = Cov(\zeta, \boldsymbol{\delta}) = Cov(\zeta, \boldsymbol{\epsilon}) = 0$,

but, with the important exception that η is not assumed to follow a normal distribution, Klein and Moosbrugger (2000) developed an adaptation of the EM algorithm. It allows to estimate the model parameters in the presence of the non-normality induced by the inclusion of the interaction between the exogenous latent variables.

Their first observation is that the exogenous latent variable can be rewritten as $\boldsymbol{\xi} = \boldsymbol{A}\boldsymbol{z}$ using the Cholesky decomposition, with $\boldsymbol{z}$ being a standard normally distributed random vector. Then, the positive definite variance-covariance matrix $\boldsymbol{\Sigma}_\xi$ of the normally distributed $\boldsymbol{\xi}$ can be rewritten as $\boldsymbol{\Sigma}_\xi = \boldsymbol{A}\boldsymbol{A}'$. This allows for estimating the lower triangular matrix $\boldsymbol{A}$ instead of $\boldsymbol{\Sigma}_\xi$ since the Cholesky decomposition is unique for positive definite matrices.

Now, the coefficient matrix $\boldsymbol{\Omega}$ can be rearranged as an upper triangular matrix (by simultaneously rearranging $\boldsymbol{\xi}$) in which the first rows contain the coefficients of those p latent variables that interact with each other to influence η and the last rows are filled with zeros if there are latent variables that don't interact with each other but only influence η linearly.

Accordingly, the vector $\boldsymbol{z}$ can be partitioned so that $\boldsymbol{z}_1$ represents those latent variables that interact with each other in the model and $\boldsymbol{z}_2$ are those latent variables that influence the endogenous variable η only linearly

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{z}_2 \end{pmatrix}$$

with $\boldsymbol{z}_1 = (z_{1,1}, \ldots, z_{1,p})$ and $\boldsymbol{z}_2 = (z_{2,p+1}, \ldots, z_{1,q})$. If these transformations are inserted into Equations (1.1) and (1.2), the model can be separated in a linear and a nonlinear part. This implies that the joint distribution of the manifest variables $(\boldsymbol{X}, \boldsymbol{Y})$ is linear in $\boldsymbol{z}_2$ but nonlinear in $\boldsymbol{z}_1$. Thus, the conditional distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ on $\boldsymbol{z}_1$ is multivariate linear and the unconditional joint distribution can be expressed as

$$f(\boldsymbol{X}, \boldsymbol{Y}) = \int_{\mathbb{R}^k} \phi_{0,\mathbb{I}_k}(\boldsymbol{z}_1)\phi_{\mu(z_1),\Sigma(z_1)}(\boldsymbol{X}, \boldsymbol{Y})d\boldsymbol{z}_1 \tag{1.3}$$

where $\phi_{0,\mathbb{I}_k}(\boldsymbol{z}_1)$ and $\phi_{\mu(z_1),\Sigma(z_1)}(\boldsymbol{X}, \boldsymbol{Y})$ are the normal densities with respective mean and covariance in its indices. Klein and Moosbrugger (2000) deduced that the mean vector and variance-covariance matrix of $\phi_{\mu(z_1),\Sigma(z_1)}(X, Y)$ depend linearly on $\boldsymbol{z}_1$, except for the mean vector $\mu_y(\boldsymbol{z}_1)$ and the variance-covariance matrix $\Sigma_{yy}(\boldsymbol{z}_1)$ of the indicators $\boldsymbol{Y}$, which depend nonlinearly on

$\boldsymbol{z}_1$ since the latent variable η influencing $\boldsymbol{Y}$ depends nonlinearly on $\boldsymbol{\xi}$. That is why the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ cannot be computed analytically and is approximated using the Gauss-Hermite quadrature (GHQ). Subsequently, the EM algorithm is adapted to estimate the model parameters of Equations (1.1) and (1.2).

A detailed introduction to the EM algorithm is given in Section 1.6, from which its application to the LMS can be deduced. For a description of the EM algorithm in the LMS approach, see Klein and Moosbrugger (2000). Klein and Muthen (2007) propose another, sightly more robust approach allowing for small deviations from the assumed normal distribution of the latent variables and the error terms and with a quicker implementation.

The later in Chapter 2 introduced estimation of a multilevel IRT model with nonlinear latent variable effects will also use the GHQ and the EM algorithm, like the LMS approach. It will differ, however, in the definition of the model and the deduction of the estimation procedure.

1.2 Item Response Theory

Item Response Theory and Classical Test Theory are both founded on the work of Charles Spearman (1904), mentioned previously. He introduced the idea of separating the true score and random error when describing an observed variable. Louis Thurstone built on Spearman's and Alfred Binet's work (van der Linden, 2016a) to further develop latent variable models. The normal ogive model by Thomson (1919) was extended by Lord (1952) to include two parameters. Shortly after, Rasch (1960) developed an alternative representation using a logistic model – the Rasch model. Lord and Novick (1968) showed that both models (normal ogive and Rasch) result in almost equal parameter estimates by introducing a scaling parameter ($D \approx 1.7$) in the Rasch model.

The logistic model was quickly further developed, and Rasch introduced a more general model that included polytomous responses (Rasch, 1961), which was finally formalized by Andersen (1977) and Andrich (1978).

Alan Birnbaum relied on the work of Lord to build his three-parameter logistic model (3PL) and his "chapters in Lord and Novick (1968) laid the statistical foundation for maximum-likelihood estimation of the item and ability parameters in this model" (van der Linden, 2016a).

The number of contributions to the field grew each year, as can be seen in the fact that the first Handbook of Item Response Theory by van der Linden and Hambleton (1997) consisted of only one volume while the most recent consists of three volumes (van der Linden, 2016a, 2016b, 2016c). In the first volume, the interested reader can find an extensive historical review of the development of IRT.

In this section, IRT as a logistic model is introduced at the example of the classical four-parameter logistic model (4PL, Barton and Lord (1981)), with the more restrictive and more widely used 1PL, 2PL, and 3PL as special cases. Subsequently, nonlinear models in the context of IRT will be discussed.

1.2.1 Introduction of the IRT Logistic Model

The literature and research on IRT models is extensive and they can therefore have many different notations and representations. To build a common basis, a widely used parameterization in the educational context will be applied where ability is seen as a realization of a latent random variable that follows a specified distribution.

Let ξ_j be the latent ability of a person $j = 1, \ldots, J$, which is usually assumed to follow a normal distribution $\xi_j \sim \mathcal{N}(0, \sigma_\xi)$. The probability of correctly solving a dichotomous item $i = 1, \ldots, I$ is expressed by the so-called 4PL model

$$P(y_{ij} = 1|\xi_j, g_i, c_i, \gamma_i, \delta_i) = g_i + (c_i - g_i)\frac{1}{1 + \exp(-(\gamma_i(\xi_j - \delta_i))}, \qquad (1.4)$$

where γ_i is the item discrimination or slope of the latent ability, which indicates how well an item can differentiate between different individuals.

The parameter δ_i is the difficulty of an item or item location. The higher its value, the less likely a individual is able to correctly solve the item. If the sign of δ_i is reversed in the argument of the exponential, it is often called easiness.

The guessing parameter g_i represents the probability of an individual guessing the correct answer.

The ceiling effect c_i stands for the probability of any individual making a careless error and answering the item wrongly[2].

If the careless error parameter c_i is set to 1, one obtains the 3PL model by Birnbaum (1968). Additional setting the guessing parameter g_i to 0 yields the 2PL model. The 1PL model – which is equivalent to the Rasch model in its mathematical representation – also has the discrimination parameters $\boldsymbol{\gamma_i}$ set to 1.

The 1PL, 2PL, and 3PL are widely used for dichotomous data, but many extensions have been constructed, which will briefly be discussed in the following paragraphs.

The conditioning on the item parameters g_i, c_i, γ_i, and δ_i in Equation (1.4) will be left out in future model descriptions to enhance the readability.

[2] Another widely used notation defines $\gamma = a, \delta = b, g = c, c = d$.

1.2.2 Multidimensional IRT and further Extensions

Masters (1982) introduced the Partial Credit Model (PCM) to incorporate answers with more than two categories. It is an extension of the Rasch model that can be applied to data with ordinal instead of dichotomous answer categories. One advantage of the PCM is its membership in the Rasch family, which means that the PCM shares the property of sufficient statistics for the item (e.g., difficulty) and person (e.g., latent ability) parameters (Andersen, 1977).

The Generalized Partial Credit Model (GPCM) by Muraki (1992) combines the PCM and the 2PL model. It includes an item discrimination parameter, which leads to a loss of the assumption of sufficient data statistics. Another general framework for ordinal data is the Graded Response Model (GRM), which was first introduced by Samejima (1969). The differences between the GRM and PCM are discussed in Masters (1982).

The inclusion of multidimensional latent variables is an important development because it allows for more realistic model descriptions (e.g., McKinley and Reckase (1983), Reckase (2009)). Mathematically, this requires a slight change in the model

$$P(y_{ij} = \frac{1}{1+\exp(-(\boldsymbol{\gamma}_i\boldsymbol{\xi}_j - \delta_i))} \tag{1.5}$$

so that $\boldsymbol{\gamma}_i$ is now an n-dimensional row vector of coefficients of the n-dimensional column vector $\boldsymbol{\xi}_j$ of the multidimensional ability.

The difficulty δ_i now has to be interpreted differently (Reckase, 2009). In a uni-dimensional IRT model, the difficulty defines the value of the latent variable at which the probability of answering an item correctly is 50%. In models with two latent variables, infinitely many combinations of both latent variables exist that result in a 50% probability of answering the item correctly. If, for example, $\gamma_1 = 1$, $\gamma_2 = 0.4$, and $\delta_i = -1$, then the intersections of the line with the coordinate axis are at $-(-1)/1 = 1$ and

$-(-1)/0.4 = 0.4$, respectively, and all combinations of latent variables on the line have a 50% probability of a correct answer (see the line in Figure 1.2).

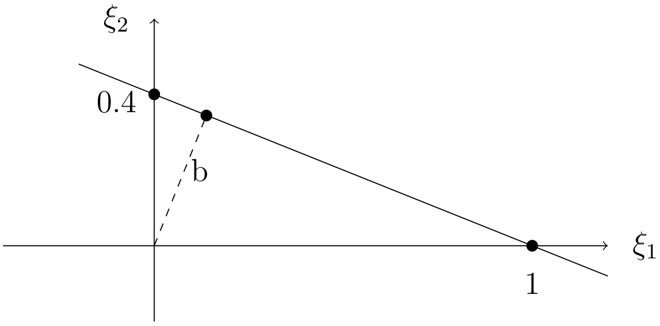

Figure 1.2: Illustration of the multivariate difficulty for an item in a two-dimensional IRT model with given parameter values. Each axis represents one latent dimension. The points on the line are those factor scores that result in a probability of 50% of answering the item correctly. The multivariate difficulty can be defined as the distance b between the line and the origin.

However, the distance between that line and the origin of the vector space is a unique indicator of the "difficulty" of the respective item (see the dashed line in Figure 1.2) in a uni-variate sense. This is often called the mdiff parameter (Reckase, 2009). It can be calculated by dividing the parameter δ_i in (1.5) by the Euclidean length of the coefficient vector of the latent variable:

$$b_i = \frac{-\delta_i}{\sqrt{\gamma_i \gamma_i'}}.$$

In the example in Figure 1.2, the mdiff for that item would be given by $b = -(-1)/\sqrt{1^2 + 0.4^2} \approx 0.9$. All this notwithstanding, the parameter δ_i will be called difficulty throughout the thesis like in the uni-dimensional case.

1.2.3 Nonlinear IRT Models

Before discussing nonlinearity in IRT models, it first needs to be defined what is meant by *nonlinear IRT models*. A logistic function, for example, is already nonlinear, which is why IRT in general is seen as a nonlinear model by scholars such as Vermunt (2004). Nonlinearity in this thesis implies a nonlinear influence of the latent variable within the exponential of the IRT function. Nonlinearity can be introduced, for example, by adding quadratic effects of a latent variable. The use of multidimensional latent abilities allows for the introduction of interactions between latent variables in the IRT model. Most publications, however, deal with nonlinear effects in frameworks other than IRT (e.g., Structural Equation Models, Generalized Linear Latent and Mixed Models), which will be discussed in later sections of this chapters.

Nonparametric IRT models (e.g., Mokken (1971)), on the other hand, implicitly also include possible nonlinear effects of the latent variable. However, in these models, effect sizes cannot be evaluated and are more of a tool for exploratory research (e.g., Sijtsma and Meijer (2007) and Sijtsma (1998)).

General Linear Latent Variable Model

Nonlinear latent effects within IRT have only been discussed by Rizopoulos and Moustaki (2008). They introduce nonlinear latent variable effects in the context of the General Linear Latent Variable Model (GLLVM) framework (Bartholomew, 1987; Bartholomew et al., 2011).

The GLLVM assumes that the correlations between manifest variables $\mathbf{X} = X_1, \ldots, X_m$ can be explained by two[3] latent variables $\boldsymbol{\xi} = (\xi_1, \xi_2)$ and observed covariates $\boldsymbol{w}$. It is further assumed that the distribution of the man-

[3] The model theoretically accounts for an undefined number of latent variables, but is restricted to two latent variables here for easier model description.

ifest variables follows an exponential family

$$P(X_i|\boldsymbol{\xi}, \boldsymbol{w}, \theta_i, \phi_i) = \exp\left(\frac{X_i\theta_i - b_i(\theta_i)}{a_i(\theta_i)} + d_i(X_i, \phi_i)\right) \tag{1.6}$$

with $i = (1, \ldots, m)$, where θ_i and ϕ_i are the natural and dispersion parameters of the exponential family. Since the functions $a_i(\theta_i)$, $b_i(\theta_i)$, and $d_i(X_i, \phi_i)$ are defined for each manifest variable individually, each manifest variable can follow a different member of the exponential family, making these models very flexible and widely applicable. Therefore, dichotomous variables are also allowed, which is necessary for IRT models.

Equation (1.6) also depends on the latent variables $\boldsymbol{\xi}$ and on the observed covariates $\boldsymbol{w}$. They are introduced to the model by including a linking function $g(\cdot)$ as in Generalized Linear Models (GLM). The manifest variables $\boldsymbol{X}$ are connected to $\boldsymbol{\xi}$ and possible covariates $\boldsymbol{w}$ by

$$g(\mathbb{E}\left[X_i|\boldsymbol{W}\right]) = \boldsymbol{W}\beta_i^{(w)} + \Xi\beta_i^{(\xi)}$$

where $\boldsymbol{W}$ and Ξ stand for the design matrices of the observed covariates and the latent variables, respectively. The matrix Ξ is built as a function of the latent variable $\boldsymbol{\xi}$, which may include nonlinear latent variable effects. Logistic IRT models can be built within this approach using the logit link function.

From this general model description, Rizopoulos and Moustaki (2008) derived the observed likelihood function $l(\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ are the model parameters. They noticed that this derivative is also the expectancy of the derivative of the complete data log-likelihood conditional on the posterior probability of the latent variable given the data. As will be seen later, the conditional expectancy of the complete data log-likelihood with respect to the posterior probability of the latent variable is the main component of the Expectation-Maximization algorithm that will be introduced in Section 1.6. The model parameters can therefore either be estimated by applying the derivative $\partial l(\boldsymbol{\omega})/\partial\boldsymbol{\omega}$ of the observed likelihood function as an EM algorithm, or by directly finding the root of $l(\boldsymbol{\omega})$. This implies that the model by Rizopoulos and Moustaki (2008) can be estimated using a hybrid algorithm. The EM is applied to quickly converge to the proximity of a solution and then a *common* optimization is carried out after a certain number of iterations, using the same objective function. This is because the biggest strength of the EM is its quick convergence at the beginning. However, since it converges more slowly the nearer it gets to the optimum, the algorithm can be switched to the *common* minimization of the derivative of the observed likelihood, which can be expected to converge quickly close to the solution.

In other implementations, hybrid algorithms that switch between the EM and a common Maximum Likelihood Estimation (MLE) have been used as well. But in these implementations, the algorithm is applied to the observed log-likelihood and the complete log-likelihood directly instead of their derivatives.

1.3 Multilevel Models

In this section, after clarifying the terminology of multilevel models, three types of multilevel models are introduced. First, the common notation is given with the multilevel regression model. Second, an extension of SEMs is presented to include a multilevel structure. Third, multilevel models in the context of IRT are discussed.

1.3.1 Terminology of Multilevel Models

SEMs were developed to analyze the relationships between latent variables. IRT evolved in the context of tests and tries to capture the influence of latent ability and item covariates on the probability of solving an item. Nevertheless, in both frameworks, data can occur that exhibits dependencies between groups and independence only within groups.

An example is provided by income. There are differences in mean income between countries, but incomes within a country are independent of each other[4].

Models that account for such clustered / dependent data were developed under many different names in different fields, such as random effects models (statistics, econometrics), linear mixed models (statistics), variance components models (statistics), hierarchical linear models (education, Bayesian), multilevel models (sociology, education), contextual effects models (sociology), random-coefficient models (econometrics), and repeated measures models / repeated measures ANOVA (statistics, psychology).

Due to this development in many different fields and frameworks, different terminology is used for the clustering, like levels or groups; the different clusters may be called Level 1, Level 2, and so on or the between- and within-cluster levels (in two-level models only).

This thesis focuses on IRT models, which are mostly used in the educational context, where the term multilevel models was coined. This term will be used here as well. For a more detailed introduction, see e.g., Snijder and Bosker (2012), Heck and Thomas (2015), and Hox, Moerbeek, and van de Shoot (2018).

1.3.2 Multilevel Regression Model

Multilevel models try to capture unobserved heterogeneity between observed variables by introducing a random variable. Effects that commonly affect all observed variables are assumed to follow a distribution, which is mimicked by that random variable. This allows for homogeneity within clusters while differences between the clusters are modeled by a latent variable. One

[4] The model could of course be extended to show dependencies within a country as well by including more clustered structures.

classical way of defining a multilevel model is a multilevel regression model

$$y_{jk} = \beta_{0k} + \beta_{1j} x_{jk} + \epsilon_{jk} \tag{1.7}$$

$$\beta_{0k} = \alpha_{00} + \alpha_{01} w_k + \delta_{0k} \tag{1.8}$$

$$\beta_{1k} = \alpha_{10} + \alpha_{11} w_k + \delta_{1k} \tag{1.9}$$

with

$$\begin{aligned} \epsilon_{jk} &\sim \mathcal{N}(0, \sigma^2) \\ \delta_{0k} &\sim \mathcal{N}(0, \sigma_0^2) \\ \delta_{1k} &\sim \mathcal{N}(0, \sigma_1^2). \end{aligned}$$

The y_{jk} is the the value of the dependent variable for individual $j = 1, \ldots, J$ in cluster $k = 1, \ldots, K$ with independent variables x_{jk} and cluster covariate w_k. The variable α_{00} is the overall intercept and α_{10} is the mean coefficient of x_{jk}. The random effect ϵ_{jk} is the individual-specific error term, δ_{0k} is the cluster-specific random effect with normal distribution and δ_{1k} is the cluster-specific random effect for the slope of x_{jk}. The random effects δ_{0k} and δ_{1k} can also be seen as error terms at the cluster level. Often, their covariances are not assumed to be zero, while the covariances between them and the individual error term ϵ_{jk} are assumed to be zero.

Substituting equations (1.7) and (1.8) into (1.9) and rearranging them yields

$$\begin{aligned} y_{jk} &= (\alpha_{00} + \alpha_{01} w_k + \delta_{0k}) + (\alpha_{10} + \alpha_{11} w_k + \delta_{1k}) x_{jk} + \epsilon_{jk} \\ \Leftrightarrow y_{jk} &= \alpha_{00} + \alpha_{01} w_k + \alpha_{10} x_{jk} + \alpha_{11} w_k x_{jk} + \delta_{0k} + (\delta_{1k} x_{jk} + \epsilon_{jk}). \end{aligned}$$

Depending on how δ_{1k} is interpreted, either the influence of the independent variable x_{jk} changes in each cluster (δ_{1k} as random effect) or the error in each cluster depends on the value of x_{jk} (δ_{1k} as error term). Additionally, an interaction term is automatically created between the independent variables on the cluster level w_k and on the individual level x_{jk}. This it not, however, an interaction between latent variables.

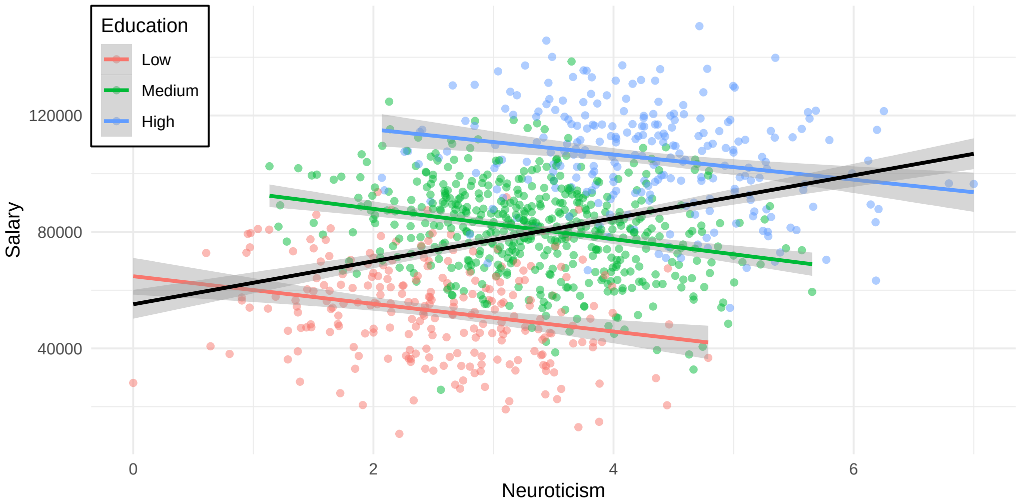

Figure 1.3: Example for multilevel data. The salary depends on neuroticism, while the education of each individual influences their relationship. This multilevel model allows for different means in each cluster (educational background), resulting in parallel regressions. In addition, the well-known Simpson's paradox is shown here, with the overall relationship (black line) inversed correlated to the relations within each cluster (colored lines). The figure is inspired by van der Laken (2017).

The multilevel regression model allows different means between groups (see Figure 1.3) as well as varying influences of explanatory variables across groups (see Figure 1.4). Figure 1.3 also depicts the well-known Simpsons paradox, which describes cases in which the overall correlation is in the opposite direction as the correlation within each cluster. On the one hand, this paradox provides a rational for applying multilevel modeling, while on the other hand, the example warns to not blindly transfer relationships from one level to another.

Multilevel regression and structural equation models were mostly developed independently. Nevertheless, over time it became clear that there is equivalence between some models in both frameworks. Curran (2003) and Bauer (2003), for example, explore how SEM can be used to estimate multilevel regression models and show its advantages and limitations. In the following subsection, multilevel Structural Equation Models will be introduced.

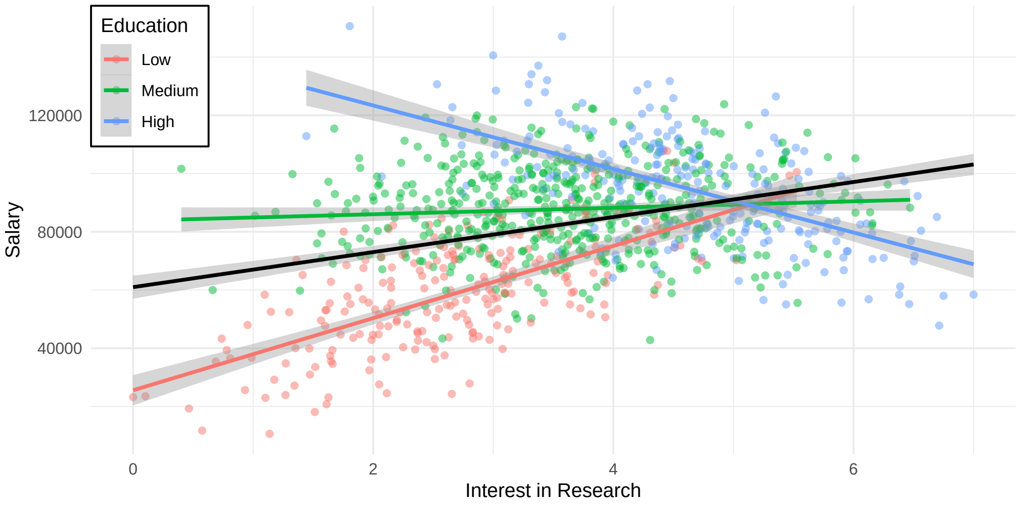

Figure 1.4: This multilevel model allows for different slopes within each cluster in addition to different means. In this simulated example, salary depends on interest in research or interest in gaining knowledge, but the relationship changes depending on the level of education. Persons with lower education might use their thirst for knowledge to gain more insight and reach higher levels of responsibility on their job. Persons with high education, however, might tend to enter careers in academia if their interest in research is high enough, which could result in a lower income than in the private sector.

1.3.3 Multilevel Structural Equation Model

There are two main ways to develop Multilevel Structural Equation Models (ML-SEMs): starting with multilevel regression or starting with structural equation models (Rabe-Hesketh, Skrondal, & Pickles, 2004). The first approach will be presented in a subsequent subsection on GLLAMM, while the second approach is introduced here.

Historical Development

“The general statistical model for multilevel SEM is complicated and, as a practical matter, was difficult to implement in software programs because of the complexities in computing separate variance-covariance matrices for units of varying sample sizes" (Heck & Thomas, 2015). However, Goldstein and McDonald (1988), McDonald and Goldstein (1989), B. Muthén (1989), and Lee (1990) almost simultaneously published approaches for a two-level SEM. Later, Liang and Bentler (2004) showed similarities between these approaches and built an EM estimator, representing a partly[5] unifying framework. Where further developments to these approaches and even alternative approaches exist, most are based on the model presented in Liang and Bentler (2004), which is why it will be used to briefly introduce the (two-level) ML-SEM. This will be done using the model built by B. Muthén (1989).

The Model

First, observations $\boldsymbol{z}_k$ $(k = 1, \ldots, K)$ on the cluster level are assumed to be independent and identically distributed and observations $\boldsymbol{y}_{jk}$ $(j = 1, \ldots, N_k)$ on the individual level are assumed to be only independent and identically distributed within a given cluster. The individual cluster sizes are denoted by N_k. Second, the individual observations are decomposed into uncorrelated random vectors $\boldsymbol{y}_{jk} = \boldsymbol{u}_k + \boldsymbol{\xi}_{jk}$ representing the between- and within-level effects, respectively. As in McDonald and Goldstein (1989) and B. Muthén (1989), it is assumed that $\boldsymbol{\xi}_{jk}$ and $\boldsymbol{z}_k$ are uncorrelated. The observed variables can therefore be summarized as

$$\begin{pmatrix} \boldsymbol{z}_k \\ \boldsymbol{y}_{jk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}_k \\ \boldsymbol{u}_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\xi}_{jk} \end{pmatrix}. \tag{1.10}$$

This basic decomposition is common to all four of the aforementioned approaches and facilitates the formulation of the necessary assumptions

$$\begin{pmatrix} \boldsymbol{z}_k \\ \boldsymbol{u}_k \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_B) \quad \text{with } \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_z \\ \boldsymbol{\mu}_u \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{pmatrix} \boldsymbol{\Sigma}_{zz} & \boldsymbol{\Sigma}_{zu} \\ \boldsymbol{\Sigma}_{uz} & \boldsymbol{\Sigma}_{uu} \end{pmatrix}$$

[5] Although these approaches are not completely mathematically equivalent, they show significant similarities.

and

$$\boldsymbol{\xi}_{jk} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_W)$$

where all variance-covariance matrices are positive definite. Following B. Muthén (1989), the measurement and structural models can then be written as

$$\begin{aligned} \boldsymbol{y}_{jk} &= \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{jk} + \boldsymbol{\epsilon}_{jk} \\ \boldsymbol{\eta}_{jk} &= \boldsymbol{\alpha}_c + \boldsymbol{\pi}_{jk} \\ \boldsymbol{\alpha}_k &= \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{z}_k + \boldsymbol{\delta}_k \end{aligned} \tag{1.11}$$

where $\boldsymbol{\nu}$ is the vector of the overall means, $\boldsymbol{\Lambda}$ is the parameter matrix, and $\boldsymbol{\epsilon}_{jk}$ is the measurement error vector with mean 0. The latent variable vector $\boldsymbol{\eta}$ in turn has a mean that is determined by the overall mean $\boldsymbol{\alpha}$, the observable Level 2 variables $\boldsymbol{z}_k$ (with coefficient matrix $\boldsymbol{\Gamma}$), a random vector $\boldsymbol{\delta}_k$ with mean 0, and $\boldsymbol{\pi}_{jk}$ which represents the random individual influence.

Now, the basic decomposition in (1.10) can be rewritten to match this specific model

$$\begin{pmatrix} \boldsymbol{z}_k \\ \boldsymbol{y}_{jk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}_k \\ \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Lambda}\boldsymbol{\pi}_{jk} + \boldsymbol{\epsilon}_{jk} \end{pmatrix} \tag{1.12}$$

and the values of the mean vectors and variance-covariance matrices in the assumptions can be updated accordingly.

Further Research

It should be noted that in this original formulation only one latent variable on the within-level is possible. It is also important to mention that this is only one representation of a ML-SEM and there are several further developments based on the approaches mentioned earlier (e.g., B. Muthén (1991), B. Muthén (1994), B. Muthén (1997), Kelava and Brandt (2009), Kelava and Brandt (2014), and Rockwood (2019)).

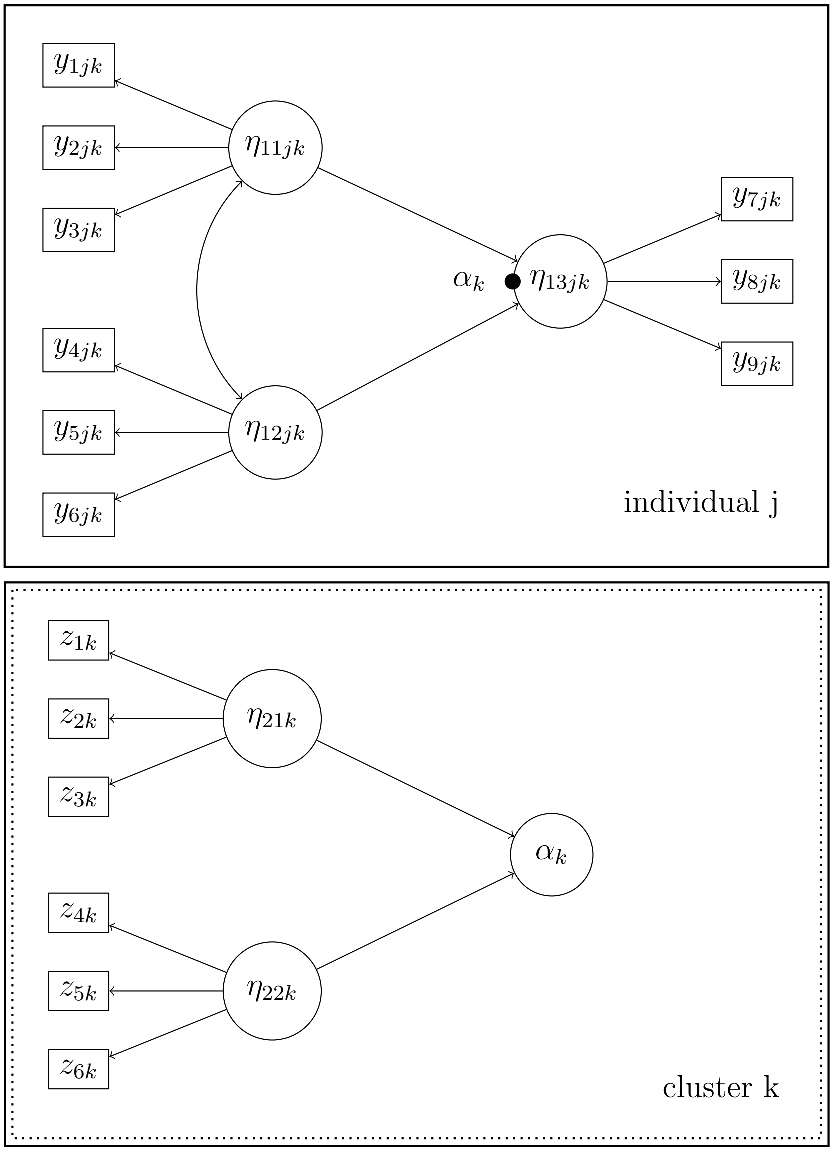

Figure 1.5: Depiction of a multilevel structural equation model. The upper rectangle shows the relationships among three latent variables on Level 1. One latent variable has a random intercept α_k whose Level 2 relations are shown in the lower rectangle.

Figure 1.5 provides an illustrative depiction of a more complex multilevel SEM from Kelava and Brandt (2014). In the upper rectangle – on Level 1 – are three latent variables η_{11jk}, η_{12jk}, and η_{13jk} that are each measured by three manifest variables y_{ijk} ($i = 1, \ldots, 9$). The affiliation to Level 1 is shown by the description ‘individual j’. The illustration is the same as a structural equation model without a hierarchical structure, but with an additional index k and a dot with the label α_k. The variable α_k stands for a random variable (as did α_k in 1.11), so that the latent variable η_{13jk} has a (potentially) different intercept in each cluster.

The variable α_k is assumed to follow a distribution. In the lower frame are further relationships with α_k that take place on Level 2. The affiliation to Level 2 is shown by the description ‘cluster k’. The figure shows that in each cluster the random intercept α_k is influenced by two Level 2 latent variables η_{21k} and η_{22k} that are each measured by 3 Level 2 manifest variables.

Today, multilevel SEM based on SEM (ML-SEM) or multilevel regression can be estimated with many different software solutions using frequentist methods, including lavaan (in R) (Rosseel, 2012), Mplus (L. Muthén & Muthén, 1998-2017), EQS (Bentler, 2006), or Bayesian methods, among which are Mplus, (Open)Bugs (Thomas, 2005), JAGS (Plummer, 2007), and Stan (Stan Development Team, 2018). For a more detailed introduction, see e.g. Heck and Thomas (2015), Lee (2007), and Mehta and Neale (2005).

1.3.4 Multilevel Item Response Theory Models

Multilevel IRT models are especially useful and necessary in educational contexts to account for clustered structures like students in classes or schools. Over the years, different approaches have been developed. Adams, Wilson, and Wu (1997) and Mislevy and Bock (1989), for example, both combined multilevel regression models with the IRT framework. Later, Fox and Glas (2001) introduced a broad model that integrated multilevel features into IRT models in the context of Bayesian estimation. As for SEM, there is also a connection between multilevel regression models and IRT. Singer (1998) and Kamata (2001), for example, showed that Rasch models can also be estimated as multilevel regression models. For a more detailed introduction, see e.g., Kamata and Vaughn (2010) or Sulis and Toland (2017).

As several, very similar, approaches exist, a multilevel IRT model will be presented that is (on paper) straightforward. A multilevel 2-PL model with random intercept can be built by adding a random intercept $u_k \sim \mathcal{N}(0, \sigma_u^2)$ that varies on the cluster level, so that

$$P(y_{ijk} = 1|\xi_{jk}, u_k) = \frac{1}{1 + \exp(-(\gamma_i \xi_{jk} - \delta_i + u_k))} \tag{1.13}$$

where the difficulty δ_i and the latent ability ξ_{jk} are defined as in Equation (1.4) above. This model has a slightly different parameterization than Fox's model, as he – like many others – adds the random intercept as the mean of the latent ability ξ_{jk}:

$$P(y_{ijk} = 1|\xi_{jk}, u_k) = \frac{1}{1 + \exp(-(\gamma_i(\xi_{jk} + u_k) - \delta_i))}. \tag{1.14}$$

The main difference between the two models lies in their interpretation. In the first model (1.13), the cluster affects the mean solving probability within each cluster, and in the second model (1.14), the cluster influences the mean ability of each student within a cluster.

A random slope can be added by letting γ_i vary on the cluster level as well by setting $\gamma_{ik} = \beta_i + u_{ik}$ where $u_{ik} \sim \mathcal{N}(0, \sigma_{\gamma_i}^2)$, which results in

$$\begin{aligned} P(y_{ijk} = 1|\xi_{jk}, u_k) &= \frac{1}{1 + \exp(-((\beta_i + u_{ik})\xi_{jk} - \delta_i + u_k))} \\ &= \frac{1}{1 + \exp(-(\beta_i \xi_{jk} + u_{ik}\xi_{jk} - \delta_i + u_k))} \end{aligned}$$

Technically speaking, this model introduces an interaction between latent variables but *not* an interaction effect of the latent abilities. The model is still linear in its predictors.

Even though this model can be built on paper, the existing software only allows either a random slope or a random intercept to be estimated.

1.4 Generalized Linear Latent and Mixed Models

As mentioned before, multilevel SEM can be built using either multilevel regression models or structural equation models as a foundation. Thus, Rabe-Hesketh et al. (2004) extended GLMs to include latent variables. This created the GLLAMM framework, which combines the previously defined constructs – SEM, IRT, and multilevel modeling. As a rational for their approach, Rabe-Hesketh et al. (2004) noted disadvantages of previous approaches, such as the potential need of different types of balances in ML-SEM (as presented above) – complete multivariate responses without missing items, a balanced multilevel design, and balanced covariates with the same sets of values on every level – and realized that they are not necessary when using multilevel regression (Skrondal & Rabe-Hesketh, 2004).

1.4.1 The General Model

The framework developed by Rabe-Hesketh et al. (2004) is divided into three parts – the response model, the structural model for the latent variables, and the distribution of the latent variables. These three parts will be presented next before focusing on more detailed descriptions and applications of the GLLAMM framework. The notations given in Rabe-Hesketh et al. (2004) will be adjusted to match the notation used in this thesis.

The Response Model The first step in their response model is to define the *linear* predictor ν for K levels and N_k latent variables. The first level is reserved for fixed effects only, so that $k = 2, \ldots, K$, which results in

$$\nu = \boldsymbol{\beta}'\boldsymbol{w} + \sum_{k=2}^{K} \sum_{n=1}^{N_k} \xi_n^{(k)} \boldsymbol{\lambda}_n^{(k)\prime} \boldsymbol{z}_n^{(k)}. \tag{1.15}$$

The elements of $\boldsymbol{w}$ are the explanatory variables or covariates, which are weighted by the regression coefficients $\boldsymbol{\beta}$. The latent variables $\xi_n^{(k)}$ for each level are multiplied by a linear combination of exploratory variables $\boldsymbol{z}_n^{(k)}$ with factor loadings $\boldsymbol{\lambda}_n^{(k)}$ (usually $\lambda_1^{(k)}$ is set to 1). In a second step, the linear predictor ν – as in generalized linear models – is linked to the conditional expectation of the response vector $\boldsymbol{y}$ given the covariates $\boldsymbol{w}$. Defining $\boldsymbol{z} = (\boldsymbol{z}^{(2)\prime}, \dots, \boldsymbol{z}^{(K)})'$ (with $\boldsymbol{z}^{(k)} = (\boldsymbol{z}_1^{(k)}, \dots, \boldsymbol{z}_{N_k}^{(k)})'$) and $\boldsymbol{\xi} = (\boldsymbol{\xi}^{(2)\prime}, \dots, \boldsymbol{\xi}^{(K)\prime})'$ (with $\boldsymbol{\xi}^{(k)} = (\xi_1^{(k)}, \dots, \eta_{N_k}^{(k)})'$) the linear predictor can be written as

$$g(\mathbb{E}\left[y|\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\xi}\right]) = \nu \tag{1.16}$$

with a link function $g(\cdot)$. In a final step, a distribution from the exponential family must be chosen upon which to build the conditional expectation. By defining different fixed parameters, link functions, and distributions in these three steps, different models can be built.

Structural Model As in SEM, the structural model is given by

$$\boldsymbol{\xi} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{B}\boldsymbol{W} + \zeta$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ are $N \times N$ and $N \times Q$ parameter matrices, respectively. The predictors of covariates of the latent variable is given as $\boldsymbol{W}$. The error is as usually defined as $\boldsymbol{\zeta}$.

As defined above, the vector $\boldsymbol{\xi}$ includes the latent variable vectors on each level. Therefore, this structural model theoretically allows for dependencies between latent variables on different levels. However, latent variables cannot be directly dependent on latent variables at a lower level, and since the model is a strictly recursive framework, the matrix $\boldsymbol{\Gamma}$ is an upper triangular matrix.

Distribution of the Latent Variables There are two variables that potentially need a specified distribution – the error term $\boldsymbol{\zeta}$ and the latent variable $\boldsymbol{\xi}$ when their independence is assumed. The distribution of $\boldsymbol{\zeta}$ should be specified if a structural model is defined within this framework. In other cases, its distribution may follow from the chosen link function. The distribution of the latent variables (within each level) can either be predefined or estimated, depending in the estimation method used.

1.4.2 IRT in GLLAMM

A formulation of the 1-PL, 2-PL, PCM and rating scale model in the GLLAMM framework can be found in Zheng and Rabe-Hesketh (2007) as well as in the gllamm syntax in Stata (StataCorp., 2019). Here, a two-level IRT model with a two-dimensional latent ability ($N_2 = 2$, $N_3 = 1$) will be presented.

As mentioned above, the first level in the GLLAMM context does not include latent variables but is reserved for item variables (in the IRT case), so that a two-level IRT model is a three level GLLAMM model ($K = 3$).

First, a multilevel IRT model in which the overall mean depends on the levels as in Equation (1.13) is shown. The response model in (1.15) is then written as

$$\begin{aligned}\operatorname{logit}\left(\mathbb{E}\left[y_{ijk}|\xi_{njk}^{(2)}, \xi_k^{(3)}\right]\right) =& \boldsymbol{\beta}'\boldsymbol{w}_i + \sum_{k=2}^{3}\sum_{n=1}^{N_k} \xi_{njk}^{(k)}\boldsymbol{\lambda}_n^{(k)\prime}\boldsymbol{z}_{ni}^{(k)} \\ =& -\beta_i + \xi_{1jk}^{(2)}\lambda_{1i}^{(2)} + \xi_{2jk}^{(2)}\lambda_{2i}^{(2)} + \xi_{1k}^{(3)} \qquad (1.17)\end{aligned}$$

with a logit link function $g(\cdot)$ as defined in (1.16), where the conditional expectation of y_{ijk} results in a probability, since y_{ijk} is dichotomous. In the second line of Equation (1.17), the change of the indices in $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_n^{(2)}$ is achieved by setting the i-th entry on the diagonal of $\boldsymbol{w}_i$ and $\boldsymbol{z}_{ni}^{(2)}$ to -1 and 1, respectively. Since these are matrices that (in this case) only carry information about which values of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_n^{(2)}$ influence the current item, $\boldsymbol{w}_i$ and $\boldsymbol{z}_{ni}^{(2)}$ can be left out when introducing the right indices.

The parameter $\lambda_n^{(3)}$ is instead set to 1 as is $z_{1i}^{(3)}$, so that $\xi_{1k}^{(3)}$ does not depend on the item. The structural model is given as the identity with the error $\boldsymbol{\zeta} = \mathbf{0}$.

In the IRT context, ξ_{1jk} and ξ_{2jk} can now be seen as the individuals' abilities with loadings λ_{1i} and λ_{2i}, which are independent of the cluster (no random slope). The random variable $\xi_{1k}^{(3)}$ on the third level is redefined as the random intercept u_k and β_i as the difficulty δ_i. Then, (1.17) can be rewritten in a classical IRT notation as in Equation (1.13):

$$\begin{aligned} & \operatorname{logit}\left(\mathbb{E}\left[y_{ijk}|\xi_{1jk},\xi_{2jk},u_k\right]\right) = -\delta_i + \xi_{1jk}\lambda_{1i} + \xi_{2jk}\lambda_{2i} + u_k \\ \Leftrightarrow \quad & P\left(y_{ijk}|\xi_{1jk},\xi_{2jk},u_k\right) = \frac{1}{1+\exp\left(-\left(\xi_{1jk}\lambda_{1i}+\xi_{2jk}\lambda_{2i}+u_k-\delta_i\right)\right)}. \end{aligned}$$

Second, a multilevel IRT model with random latent variable intercepts as in Equation (1.14) is shown. The GLLAMM model has again three levels, however the third level is introduced differently. The response model in (1.15) is written as

$$\begin{aligned} \operatorname{logit}\left(\mathbb{E}\left(y_{ijk}|\xi_{njk}^{(2)},\xi_{k}^{(3)}\right)\right) =& \boldsymbol{\beta}'\boldsymbol{w}_i + \sum_{k=2}^{2}\sum_{k=1}^{N_k}\xi_{njk}^{(k)}\boldsymbol{\lambda}_n^{(k)\prime}\boldsymbol{z}_{ni}^{(k)} \\ =& -\beta_i + \xi_{1jk}^{(2)}\lambda_{1i}^{(2)} + \xi_{2jk}^{(2)}\lambda_{2i}^{(2)} \end{aligned} \tag{1.18}$$

with $N_2 = 2$, a logit link function $g(\cdot)$, and $\boldsymbol{w}_i$ and $\boldsymbol{z}_{ni}^{(2)}$ defined as for Equation (1.17). The third level is introduced by defining the structural model as

$$\xi_{njk}^{(2)} = \xi_{nk}^{(3)} + \zeta_{njk}^{(2)} \tag{1.19}$$

with $N_3 = 2$ where the parameter matrix $\boldsymbol{B}$ is set to $\mathbf{0}$ and the matrix $\boldsymbol{\Gamma}$ to $\boldsymbol{I}_2$. Substituting (1.19) into (1.18) results in

$$\operatorname{logit}\left(\mathbb{E}\left[y_{ijk}|\xi_{njk}^{(2)},\xi_{nk}^{(3)}\right]\right) = \beta_i + \left(\xi_{1k}^{(3)} + \zeta_{1jk}^{(2)}\right)\lambda_{1i}^{(2)} + \left(\xi_{2k}^{(3)} + \zeta_{2jk}^{(2)}\right)\lambda_{2i}^{(2)}. \tag{1.20}$$

The redefinition of the parameters $\boldsymbol{\lambda}_{ni}^{(2)}$ (for $n = 1, 2$) and $\boldsymbol{\beta}_i$ to match the IRT context is done as before. The random variables, however, have to be defined differently. The Level 2 error terms $\zeta_{1jk}^{(2)}$ and $\zeta_{2jk}^{(2)}$ now can be seen as the variation between individuals and are redefined as the individuals' abilities ξ_{1jk} and ξ_{2jk}, respectively. The random intercepts u_{1k} and u_{2k}, on the other hand, are the re-definitions of $\xi_{1k}^{(3)}$ and $\xi_{2k}^{(3)}$, respectively.

Then, (1.20) can be rewritten in a classical IRT notation as in Equation (1.14):

$$\begin{aligned}
&\operatorname{logit}\left(\mathbb{E}\left[y_{ijk}|\xi_{1jk}, \xi_{2jk}, u_{1k}, u_{2k}\right]\right) \\
&= -\delta_i + (\xi_{1jk} + \sigma_{1k})\lambda_{1i} + (\xi_{2jk} + \sigma_{2k})\lambda_{2i} \\
\Leftrightarrow &P\left(y_{ijk}|\xi_{1jk}, \xi_{2jk}, u_{1k}, u_{2k}\right) \\
&= \frac{1}{1 + \exp\left(-\left((\xi_{1jk} + u_{1k})\lambda_{1i} + (\xi_{2jk} + u_{2k})\lambda_{2i} - \delta_i\right)\right)}
\end{aligned}$$

where δ_i is the difficulty of an item and λ_{ni} (for $n = 1, 2$) are the discrimination parameters. This model is depicted in Figure 1.6 from the point of view of SEM.

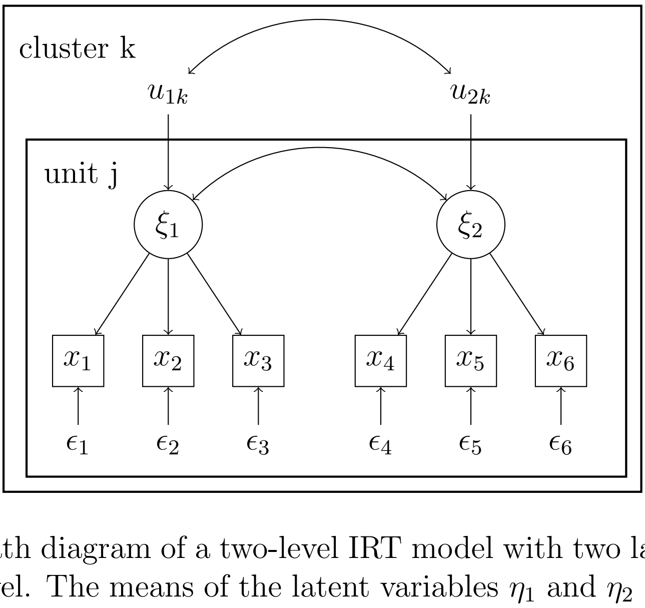

Figure 1.6: Path diagram of a two-level IRT model with two latent variables at the unit level. The means of the latent variables η_1 and η_2 depend on the random intercepts u_{1k} and u_{2k} at the cluster level, respectively, which are assumed to be correlated.

Note that within the GLLAMM framework it is not possible to build an IRT model that includes nonlinear latent variable effects, but as was shown here, other IRT models can be described to a certain extent.

1.4.3 IRT in SEM

As described above, the GLLAMM approach is a combination of SEM and GLM. It also possible to estimate some IRT models within SEM alone (e.g., Takane and de Leeuw (1987)), which will now be demonstrated using the example of the 2PL model

$$P(y_{ij} = 1|\xi_j) = \frac{1}{1 + \exp(-(\gamma_i(\xi_j - \delta_i))} \tag{1.21}$$

as defined in Equation (1.4), with ξ_j being the latent ability of a person $j = 1, \ldots, J$, γ_i describing the item discrimination, and δ_i the difficulty of an item. The measurement model is defined as in Equation (1.1)

$$\boldsymbol{X} = \boldsymbol{\Lambda}\xi + \boldsymbol{\epsilon} \tag{1.22}$$

with the manifest variables being denoted by $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)'$, the latent variable by ξ with $\boldsymbol{\Lambda}$ as its factor loading matrix, and $\boldsymbol{\epsilon}$ as the vector of measurement errors. A structural model needs to be specified along with a vector of intercepts $\boldsymbol{\tau}$. The latent variable ξ is set to be one-dimensional.

To show the equivalence of the 2PL model and the corresponding SEM measurement model, first, dichotomous items must be included in the SEM framework. In order to do so, Christoffersson (1975) assumed that the dichotomous observed variable y_{ij} actually follows an underlying logistic distribution and can be described as a metric variable y^*_{ij} with distribution parameters $\alpha = 0$ and $\beta = \sigma^2$. A threshold τ_i (not denoting the intercept of the measurement model in this context) is estimated for every item, which determines when y^*_{ij} takes the value 0 or 1[6].

[6] In the case of polytomous items, this idea is extended to include more thresholds.

Second, the relationship between the parameters in both approaches needs to be clarified. Since different parameterizations can be chosen in both approaches, here the focus is on models that are identified by fixing one factor loading to 1 and the parameter σ^2 of the underlying logistically distributed observed variable y^*_{ij} to 1.

For a detailed description of the conversions for different parameterizations, see Kamata and Bauer (2008). In essence, the matrix of loadings $\mathbf{\Lambda}$ in (1.22) corresponds to the vector of discriminations $\boldsymbol{\gamma}$ in (1.21) and the thresholds τ_{ij} correspond to the difficulties $\boldsymbol{\delta}$ in (1.21). With the chosen parameterization, the following transformations result in the parameters in IRT notation:

$$\begin{aligned}\gamma_i &= \frac{\lambda_i}{\sqrt{1-\lambda_i^2}}\\ \delta_i &= \frac{\tau_i}{\lambda_i}.\end{aligned}$$

If the 2PL model is written to account for multidimensional latent variables as in (1.5) with the exponent $\gamma_i \xi_j - \delta^*_i$, the difficulty δ^*_i is instead given by $\frac{\tau_i}{\sqrt{1-\lambda_i^2}}$.

A more intuitive way would be to assume that $y^*_{ij} \sim \mathcal{N}(0, \sigma^2)$, but that would correspond to a probit link (in GLM) and would result in the 2PL ogive model instead. However, the 2PL IRT model and the 2PL normal ogive model are closely related. The 2PL IRT model can approximate the normal ogive model by multiplying the exponent by 1.702 (see e.g., Camilli (1994)).

In both cases, estimation is usually done using MLE, which assumes normality of the continuous underlying y^*_{ij}. A more robust approach was developed by B. Muthén (1992) and B. Muthén (1997) based on Weighted Least Square (WLS) estimation that circumvents the normality assumption of y^*_{ij}. The approach is called Weighted Least Square Mean and Variance adjusted (WLSMV) estimator in `Mplus` (L. Muthén & Muthén, 1998-2017) and is also implemented in `lavaan` (Rosseel, 2012). It will be discussed in Section 3.1.1.

1.5 Estimation Techniques

In the previous sections, several models and approaches to include latent variables were introduced. The variety of estimators for these models is immense. Some of these will be briefly introduced here, with a focus on appropriate estimators for IRT models. Broader introductions to estimation procedures can be found in Bollen (1989) and Reise and Reviecki (2015).

1.5.1 Joint Maximum Likelihood Estimation

Maximum Likelihood Estimation is one of the most frequently used estimation procedures in all disciplines. MLE procedures estimate parameters of a probability distribution by maximizing a likelihood in which observed data is most probable under the assumption of a specific model.

One such MLE approach is Joint Maximum Likelihood (JML) estimation, which "is one of the earliest approaches to fitting item response theory (IRT) models. This procedure treats both the item and person parameters as unknown but fixed model parameters and estimates them simultaneously by solving an optimization problem. However, the JML estimator is known to be asymptotically inconsistent for many IRT models, when the sample size goes to infinity and the number of items keeps fixed. Consequently, in the psychometrics literature, this estimator is less preferred to the MML estimator" (Y. Chen, Li, & Zhang, 2017). Y. Chen et al. (2017) go on to re-investigate the performance of JML in high-dimensional exploratory item factor analysis and show that their adjusted estimator performs similarly to Marginal Maximum Likelihood (MML) estimators when the dimensionality of the latent variable is high enough.

1.5.2 Marginal Maximum Likelihood Estimation

MML is another approach to estimate models, in which latent variables are present – as in SEM or IRT. As the name suggests the latent variables are marginalized out of the likelihood by integrating over them, given their assumed distribution. Assuming a 2PL model as in Equation (1.4), the likelihood – usually in logarithmic form – can be written as

$$\begin{aligned}&\log L(\boldsymbol{\gamma}, \boldsymbol{\delta}|Y, \boldsymbol{\xi})\\ =&\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij}\log P(y_{ij}=1|\boldsymbol{\gamma}, \boldsymbol{\delta}, \xi_j) + (1-y_{ij})\log P(y_{ij}=0|\boldsymbol{\gamma}, \boldsymbol{\delta}, \xi_j).\end{aligned}$$

In many cases, a solution can be found by deriving the log-likelihood and finding its roots for the parameters. To achieve maximum likelihood solutions more efficiently when missing data are present Dempster, Laird, and Rubin (1977) formulated the EM algorithm based on prior developments (e.g., Hartley (1958), Brown (1974), and T. Chen and Fienberg (1976)). They showed that latent variables and parameters could be seen as missing data in this context, which led to a variety of research on the EM. Today, the EM is a frequently used maximum likelihood solution and it will be introduced in more detail in Section 1.6

1.5.3 Weighted Least Square Estimation

One possible disadvantage of likelihood approaches is the assumption of a distribution (usually, normality is assumed – especially in the social sciences) of the latent variables and – in the SEM context – of the underlying distribution of the dichotomous observed variables. B. Muthén (1984) developed a Weighted Least Square estimator that minimizes the squared difference between the observed variables and the predicted values given by the model. The weight matrix is a positive definite matrix which is the estimated asymptotic variance-covariance matrix of the polychoric correlation and threshold estimates. However, normality is still assumed in this approach. To be robust against violations of the normality assumption and thus overcome the limitations of MLE and the full WLS approach, when dichotomous variables are used, a robust weighted least square estimation was developed by B. Muthén (1992). In this approach, the weight function is only diagonal and does not need to be positive definite, which improves estimation compared to the WLS (Li, 2015).

1.5.4 Markov Chain Monte Carlo Methods

A different approach to the aforementioned estimation methods can be devised within the framework of Bayesian statistics. In Bayesian analysis, prior information about the parameters $f(\boldsymbol{\omega})$ from e.g., earlier experiments or a theoretical assessment, as well as the likelihood of the data given the parameters $f(\boldsymbol{Y}|\boldsymbol{\omega})$ are used to gain knowledge about the posterior density of the parameters $f(\boldsymbol{\omega}|\boldsymbol{Y})$ given the data. The relationship between the three densities can be described as

$$f(\boldsymbol{\omega}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\omega})f(\boldsymbol{\omega}). \tag{1.23}$$

A frequently applied method are Markov Chain Monte Carlo (MCMC) algorithms. “MCMC uses the proportionality in Equation (1.23) to evaluate the relative likelihoods of parameter estimates. Ultimately, the goal of MCMC is to reproduce the $f(\boldsymbol{\omega}|\boldsymbol{Y})$ distribution, which often cannot be determined analytically. Therefore, the characteristics of the distributions are determined by sampling enough observations from the posterior” (Natesan, Nandakumar, Minka, & Rubright, 2016).

The basic idea of MCMC is to construct a Markov chain[7] whose stationary distribution[8] corresponds to the desired distribution. A random walk

[7] A Markov chain simulates a sequence of events or states. The probability of each event depends only on the previous event(s). The shift from one neighboring event to another is called a step.

[8] A vector giving the probabilities of ending a sequence of steps at any of the events

through the Markov chain with a fixed number of steps is simulated with a given number of repetitions. Depending on the number of steps and repetitions, the probabilities of ending up in each state converge towards the stationary / desired distribution.

MCMC algorithms can differ, for example, in the way the Markov chain is built. Gibbs sampling algorithms (Geman & Geman, 1984), for instance, create a chain in which the probability of the next sample is calculated as being conditional on the current sample, whereas Metropolis-Hastings algorithms can be used when the probability of the next sample cannot be calculated directly – as Metropolis-Hastings is not conditional on the ability to sample from the posterior probability. The number of different application of MCMC methods to approximate an object is extensive and cannot be covered in its entirety here.

The Gibbs sampler can also be seen as a bayesian alternative to the EM algorithm. As will be discussed in the next section, the EM encompasses two steps – the expectation step and the maximization step – as does the Gibbs sampler, with the difference that the Gibbs sampler does not maximize over the posterior distributions (as the EM does) but rather samples datasets from the posterior distribution to estimate the parameters.

A more detailed introduction can be found in e.g., Lee (2007) or Lambert (2018).

1.5.5 Metropolis-Hastings Robbins-Monro Algorithm

Another method to obtain ML estimates is the Metropolis-Hastings Robbins-Monro (MHRM) algorithm Cai (2008). Like the EM algorithm, it is based on Fisher's Identity, which posits a connection between the observed data

in the Markov chain is called a stationary distribution if multiplying this vector with the matrix representing the transition probabilities of each event to a neighboring event results again in the same vector.

log-likelihood and the complete data log-likelihood:

$$\frac{\partial \log P(y_j|\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} = \int \frac{\partial P(y_j, \boldsymbol{\xi}|\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} P(\boldsymbol{\xi}|y_j, \boldsymbol{\omega}) d\boldsymbol{\xi} \tag{1.24}$$

where y_i is the observed variable, $\boldsymbol{\xi}$ is the latent variable and $\boldsymbol{\omega}$ is the parameter vector. Fisher's Identity will be discussed in more detail in Appendix A.1.

The MHRM uses the MCMC method "Metropolis-Hastings" to draw values for the latent variables from the posterior distribution. Subsequently, these values are used to approximate the complete data gradient as given in Fisher's identity (1.24). Finally, the parameter estimates are updated in each iteration using a so-called Robbins-Monro update, finalizing the name of the approach. A comprehensive introduction to the MHRM is given in Cai and Thissen (2015).

1.6 The Expectation-Maximization Algorithm

In the following section, the EM will be introduced in more detail, as it will subsequently be used as basis to derive the estimation procedure for a multilevel IRT model with nonlinear latent variable effects. It will be reasoned why the EM was chosen over other possible estimation approaches based on the current literature. The EM will be introduced in general terms, followed by a technical description.

1.6.1 Comparison of the EM Algorithm to Other Approaches

Naturally, many studies have compared estimation methods with different latent variable models in different circumstances. Arguably, the three most established estimation methods are the WLSMV for SEM models with ordinal data and the two MLE approaches MHRM and EM in the context of

IRT models. Han and Paek (2014) compared implementations of the MHRM algorithm and of the EM in the software Mplus (L. Muthén & Muthén, 1998-2017), flexMIRT Cai (2017), and IRTPRO (Cai, Thissen, & du Toit, 2015). They observed that all methods recovered the parameters of multivariate two-parameter IRT model with little bias. Kuo and Sheng (2016) analyzed the MHRM and an EM implementation from IRTPRO. They could replicate the finding that both yielded a similarly low bias when recovering the parameters in a graded response IRT model.

Cai (2010) compared the EM algorithm and his MHRM, finding that they achieve almost equal results but that the MHRM becomes more efficient as more dimensions of the latent variable are considered in the model.

Nevertheless, the studies by Cai (2010), Han and Paek (2014), and Kuo and Sheng (2016) might suggest that the EM with Gauss-Hermite quadrature has a slightly lower bias than the MHRM implementation, especially in small datasets and as the correlation between the latent variables $\boldsymbol{\xi}$ rises.

Sims (2017) conducted a simulation study comparing MHRM, EM, and WLSMV implemented in Mplus and IRTPRO. She could show that all three methods recovered difficulty parameters and loadings of several multidimensional IRT models well, even though the variance of the EM estimates was not stable with increasing sample size. Li (2015) found that ML could better recover interfactor correlations than the WLSMV.

Overall, EM, MHRM, and WLSMV have been shown to perform very similarly in recovering IRT parameters. Although the WLSMV performs equally well as EM and MHRM, it was developed in the context of SEM, which is why it was not chosen as a foundation for this thesis. Instead, the EM was chosen over the MHRM even though both would be valid options. Further explanations are given in Section 2.2.

1.6.2 General Introduction to the EM Algorithm

The term "EM algorithm" was first introduced by Dempster et al. (1977), who summarized the already existing literature (e.g., Hartley (1958), Brown (1974), and T. Chen and Fienberg (1976)) and provided a unifying framework resulting in more interest and research on the approach. The EM became a standard tool for problems involving missing data, which are interpreted very broadly in the EM context. Parameters and latent variables, for example, can be described as missing data. This stems from the property that – mathematically – missing data and latent variables are formulated as random variables that follow a certain distribution, which in turn can be informed by the observed data. The EM exploits the interpretation as random variable, resulting in some computational advantages.

In the context of IRT, as in model (1.5), the data can be divided into the observed data y and the unobserved ability ξ. The latter can be seen as missing values for which only the distribution is assumed – often given as $\mathcal{N}(0,1)$. Taking the information of the observed data into account, the expectation from the ability conditional on the observed data – the posterior probability – is developed. The posterior probability could be used to calculate factor scores[9] for each individual, and treat them as observed variables in the estimation of the model parameters. Instead of determining the factor scores, however, the expectation of the complete data (y, ξ) likelihood conditional on the posterior probability of the ability is analyzed in the EM. Based on this expected likelihood, the item parameters are estimated using maximum likelihood estimation.

This procedure will yield the same results as a maximum likelihood estimation based on the observed likelihood. However, one advantage of the EM in the IRT context is that it takes further information about the latent variables into account by considering the posterior probability. Another ad-

[9] Shortly put, factor scores are values that represent relative spacing of an object / individual on the latent factor. They can be seen as realizations of the random variable / latent variable.

vantage is that the conditional complete data likelihood is computationally easier to handle than the observed likelihood. In general, the EM provides a framework for easily handling missing data. For models whose density is from the exponential family, the EM further facilitates the estimation by reducing the needed data to the data's sufficient statistic (Bock & Aitkin, 1981; Dempster et al., 1977; Schilling & Bock, 2005).

1.6.3 Technical Introduction to the EM Algorithm

The EM algorithm is a framework providing tools to obtain MML estimates. It needs to be deduced for each model individually. In this section, the general idea of the EM is technically introduced, partly following Dempster et al. (1977).

Initial Definitions

Let Y be the observed data from a sample space $\mathcal{Y}$ and $\boldsymbol{\xi}$ the missing data from a sample space Ξ, so that $\boldsymbol{X} = (\boldsymbol{Y}, \xi)$ is the complete data from the sample space $(\mathcal{Y}, \Xi)$. The unobserved data can, for example, describe latent variables (as in factor scores estimation (e.g., Bartlett (1935), Thurstone (1935)), parameters of a distribution (e.g., the variance and mean of a normal distribution) or parameters of a model (e.g., item difficulty, item discrimination and item guessing parameters (e.g., Bock and Aitkin (1981), Harwell, Baker, and Zwarts (1988)).

In this thesis, latent variables $\boldsymbol{\xi} \in \mathbb{R}^n$, e.g., latent abilities of test-takers, are present in the model, which can be interpreted as missing data with a distribution $f_\xi(\boldsymbol{\xi}|\boldsymbol{\omega}^*)$ which depends on distribution parameters $\boldsymbol{\omega}^*$. The unobserved latent variables influence the observed data $y \in 0, 1^{J \times I}$, e.g., responses to test items. Here J is the number of test-taker, n is the number of latent abilities and I is the number of items administered in the test. The complete data $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{\xi})$ are assumed to follow a distributional family

$f(\boldsymbol{X}|\boldsymbol{\omega})$ depending on parameters $\boldsymbol{\omega}$ (including $\boldsymbol{\omega}^*$) in the parameter space Ω[10]. The observed data are defined to be drawn from a distributional family $g(\boldsymbol{Y}|\boldsymbol{\omega})$. The observed (incomplete) data distribution can then be written as

$$g(\boldsymbol{Y}|\boldsymbol{\omega}) = \int_{\mathbb{R}} f(\boldsymbol{X}|\boldsymbol{\omega}) d\boldsymbol{\xi} = \int_{\mathbb{R}} g(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{\omega}) f(\boldsymbol{\xi}|\boldsymbol{\omega}) d\boldsymbol{\xi}. \tag{1.25}$$

The Objective Function

As in maximum likelihood estimation, the goal is to choose those $\boldsymbol{\omega}$ that maximize the observed likelihood, which is usually written in logarithmic form

$$L(\boldsymbol{\omega}) = \log g(\boldsymbol{Y}|\boldsymbol{\omega}).$$

Because direct optimization can be complicated, however, Dempster et al. (1977) introduced an alternative. First, they introduce the posterior probability of the missing data, given the observed data:

$$k(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega}) = \frac{f(\boldsymbol{\xi}, \boldsymbol{Y}|\boldsymbol{\omega})}{g(\boldsymbol{Y}|\boldsymbol{\omega})}$$

so that the observed likelihood can be rewritten as

$$L(\boldsymbol{\omega}) = \log(f(\boldsymbol{\xi}, \boldsymbol{Y}|\boldsymbol{\omega})) - \log(k(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega})). \tag{1.26}$$

The expectations of the complete data log-likelihood $f(\boldsymbol{\xi}, \boldsymbol{Y}|\boldsymbol{\omega})$ and of the posterior probability $k(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega})$, each conditional on the observed data $\boldsymbol{Y}$ and on a different realization of the parameters $\boldsymbol{\omega}'$, can be defined as

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\omega}, \boldsymbol{\omega}') &= \mathbb{E}\left[\log(f(\boldsymbol{\xi}, \boldsymbol{Y}|\boldsymbol{\omega}))|\boldsymbol{Y}, \boldsymbol{\omega}'\right] & (1.27) \\ \mathcal{H}(\boldsymbol{\omega}, \boldsymbol{\omega}') &= \mathbb{E}\left[\log(k(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega}))|\boldsymbol{Y}, \boldsymbol{\omega}'\right]. \end{aligned}$$

[10] Not to be mistaken for the (bold) coefficient matrix $\boldsymbol{\Omega}$ of the nonlinear latent variable effects. The correct interpretation will always be clear in the context.

Building the expectancy on both sides of Equation (1.26), again conditional on the observed data $\boldsymbol{Y}$ and on a different realization of the parameters $\boldsymbol{\omega}'$, then results in

$$\begin{aligned}
&\mathbb{E}\left[L(\boldsymbol{\omega})|\boldsymbol{Y},\boldsymbol{\omega}'\right] &&= \mathbb{E}\left[\log(f(\boldsymbol{\xi},\boldsymbol{Y}|\boldsymbol{\omega})) - \log(k(\boldsymbol{\xi}|\boldsymbol{Y},\boldsymbol{\omega}))|\boldsymbol{Y},\boldsymbol{\omega}'\right] \\
\iff &\mathbb{E}\left[\log(g(\boldsymbol{Y}|\boldsymbol{\omega}))|\boldsymbol{Y},\boldsymbol{\omega}'\right] &&= \mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}') - \mathcal{H}(\boldsymbol{\omega},\boldsymbol{\omega}') \\
\iff &\log(g(\boldsymbol{Y}|\boldsymbol{\omega}')) = L(\boldsymbol{\omega}') &&= \mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}') - \mathcal{H}(\boldsymbol{\omega},\boldsymbol{\omega}') \qquad (1.28)
\end{aligned}$$

Optimization of the Complete Data Likelihood

The heuristic idea of the EM is to maximize $\mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}')$ instead of $L(\boldsymbol{\omega})$ in an iterative process. Let $\boldsymbol{\omega}^{(p)}$ be the estimates of the current iteration p. First, the function $\mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}^{(p)})$ is computed for fixed $\boldsymbol{\omega}^{(p)}$, which results in a function that depends only on the unknown $\boldsymbol{\omega}$. In an application that could mean that those terms in $\mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}^{(p)})$ are determined that depend only on the current estimates $\boldsymbol{\omega}^{(p)}$. One example could be the calculation of the posterior probability of the missing data $\boldsymbol{\xi}$ given the observed data $\boldsymbol{Y}$ and the current estimates of the parameters $\boldsymbol{\omega}^{(p)}$:

$$\begin{aligned}
P(\boldsymbol{\xi}|\boldsymbol{Y},\boldsymbol{\omega}^{(p)}) &= \frac{f(\boldsymbol{\xi},\boldsymbol{Y}|\boldsymbol{\omega}^{(p)})}{g(\boldsymbol{Y}|\boldsymbol{\omega}^{(p)})} \\
&= \frac{g(\boldsymbol{Y}|\boldsymbol{\xi},\boldsymbol{\omega}^{(p)})P(\boldsymbol{\xi}|\boldsymbol{\omega}^{(p)})}{\int g(\boldsymbol{Y}|\boldsymbol{\xi},\boldsymbol{\omega}^{(p)})P(\boldsymbol{\xi}|\boldsymbol{\omega}^{(p)})d\boldsymbol{\xi}}.
\end{aligned}$$

In fact, the function $\mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}^{(p)})$ in (1.27) is integrated with respect to this posterior probability.

Second, $\mathcal{Q}(\boldsymbol{\omega},\boldsymbol{\omega}^{(p)})$ is maximized to find

$$\boldsymbol{\omega}^{(p+1)} = \arg\max_{\boldsymbol{\omega}} \mathbb{E}\left[\log(f(\boldsymbol{\xi},\boldsymbol{Y}|\boldsymbol{\omega}))|\boldsymbol{Y},\boldsymbol{\omega}^{(p)}\right].$$

For the expectations of the the posterior probability $k(\boldsymbol{\xi}|\boldsymbol{Y},\boldsymbol{\omega})$ conditional on the observed data $\boldsymbol{Y}$ and on a different realization of the parameters $\boldsymbol{\omega}'$

$$\mathcal{H}(\boldsymbol{\omega}, \boldsymbol{\omega}') \leq \mathcal{H}(\boldsymbol{\omega}, \boldsymbol{\omega}) \tag{1.29}$$

holds for any pair of $(\boldsymbol{\omega}, \boldsymbol{\omega}')$. The proof will follow in Section 2.3.3 in Equation (2.14). Applying (1.29), it can be proven that the maximization of $\mathcal{Q}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)})$ also maximizes $L(\boldsymbol{\omega})$:

$$\begin{aligned} L(\boldsymbol{\omega}^{(p)}) &= \mathcal{Q}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) - \mathcal{H}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) \\ &\leq \mathcal{Q}(\boldsymbol{\omega}^{(p+1)}, \boldsymbol{\omega}^{(p)}) - \mathcal{H}(\boldsymbol{\omega}^{(p+1)}, \boldsymbol{\omega}^{(p)}) \\ &= L(\boldsymbol{\omega}^{(p+1)}). \end{aligned}$$

In summary, the EM has two steps

1. *Expectation-step*: Calculate the conditional expectation $\mathcal{Q}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)})$
2. *Maximization-step*: Find $\boldsymbol{\omega}^{(p+1)} = \arg\max_{\boldsymbol{\omega}} \mathbb{E}\left[\log(f(\boldsymbol{\xi}, \boldsymbol{Y}|\boldsymbol{\omega}))|\boldsymbol{Y}, \boldsymbol{\omega}^{(p)}\right]$

that are iterated until convergence occurs. Naturally, starting values $\boldsymbol{\omega}^{(0)}$ need to be chosen before the first iteration.

Chapter 2

A Multilevel IRT Model with Nonlinear Latent Effects

In the previous chapter, latent variable models were introduced as well as extensions to multivariate latent variables, nonlinear latent variable effects, and multilevel structures. Estimation methods that allow to attain the unknown parameters were discussed and different latent variable frameworks containing some of these extensions were reviewed. Unfortunately, there seems to be no framework capable yet of simultaneously estimating an IRT model containing nonlinear latent variable effects and a multilevel structure.

In this chapter, these approaches are merged to first define the **M**ultilevel **I**RT with **No**nlinear **L**atent variable **E**ffects **M**odel (**MINoLEM**). An estimation procedure for this model will be presented by deriving the likelihood function of an EM algorithm and its derivative. Necessary implementation steps will be established, including the search for appropriate starting values for the optimization.

2.1 Definition of the Model

The model is based on the 2PL IRT model with multivariate latent variables within the IRT framework (1.5). Alternatively, it can also be seen as an extension of the GLLAMM framework to include nonlinear latent variable effects in the *linear* predictor in Equation (1.15).

A multilevel structure is considered by adding a random intercept u_k that follows a normal distribution $\mathcal{N}(0, \sigma^2)$. This random intercept changes the complete exponent by a different constant in every cluster rather than changing the mean abilities in every cluster as in some multilevel models (see e.g., Fox's model in Equation (1.14)). This also influences the definition of the ability $\boldsymbol{\xi}_{jk}$, which is a multidimensional $(n \times 1)$ vector of person parameters or latent abilities (e.g., reading ability, math ability, listening ability). These belong to a person j $(j = 1, ..., J_k)$ from a cluster k $(k = 1, ..., K)$, and follow a multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_\xi)$.

Finally, an interaction / quadratic effect of the latent variable is included in the model by adding $\boldsymbol{\xi}'_{jk}\boldsymbol{\Omega}_{ik}\boldsymbol{\xi}_{jk}$ into the exponent. The parameter matrix $\boldsymbol{\Omega}_i$ is a $(n \times n)$ lower triangular matrix indicating the loadings for every interaction / quadratic effect of the latent ability $\boldsymbol{\xi}_{jk}$.

The probability of answering a dichotomous item correctly is then given by

$$P(Y_{ijk} = 1|\boldsymbol{\omega}, \boldsymbol{\xi}_{jk}, u_k) = \frac{1}{1 + \exp(-(\boldsymbol{\xi}'_{jk}\boldsymbol{\gamma}_i + \boldsymbol{\xi}'_{jk}\boldsymbol{\Omega}_i\boldsymbol{\xi}_{jk} - \delta_i + u_k))} \tag{2.1}$$

with

$$\boldsymbol{\omega} = (\gamma_1, \ldots, \gamma_I, \Omega_1^{11}, \ldots, \Omega_1^{nn}, \ldots, \Omega_I^{11}, \ldots, \Omega_I^{nn}, \delta_1, \ldots, \delta_I, \boldsymbol{\Sigma}_\xi, \sigma^2). \tag{2.2}$$

The parameter vector $\boldsymbol{\omega}$ includes the difficulties δ_i for every item, the coefficients $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Omega}_i$ of the latent variables for every item,the variance-covariance matrix $\boldsymbol{\Sigma}_\xi$ of the latent variables, and the variance σ^2 of the random intercept u_k.

2.2 The Choice of Framework and Estimation Method

In this section, the choice of the EM algorithm as estimation method will be explained. First, the choice of IRT as framework will be discussed.

The model (2.1) could also have been framed within the SEM context. However, this would have required essentially three adjustments to the general model – dichotomous items, nonlinear latent variable effects, and a multilevel structure. Therefore, it seemed more straightforward to circumvent the transformation of continuous variables to binary variables and choose IRT directly.

Nevertheless, the LMS approach by Klein and Moosbrugger (2000) gave some insights into handling nonlinearity. They tackled the well-known nonlinear model by Kenny and Judd (1984) and added possible interactions and quadratic effects to their model. Furthermore and more importantly, they estimated the model using an EM algorithm. The advantage of framing the problem with IRT instead of SEM is that the nonnormal distribution resulting from latent interactions or quadratic effects does not need to be taken into account as it does in LMS. It is possible to directly use numerical integration if a normal distribution is assumed for each individual latent variable.

Additional understanding stems from the GLLAMM by Rabe-Hesketh et al. (2004), whose very broad framework already incorporates dichotomous data and multilevel structures. However, possible models do not include nonlinear latent variable effects. Rizopoulos and Moustaki (2008) introduced nonlinear latent variable effects within the similar GLLVM framework by Bartholomew et al. (2011). They base their approach on models that can be represented by the exponential family, which is, unfortunately, not straightforward for the MINoLEM. The more promising approach is the adoption of IRT and instead being inspired by GLLAMM, GLLVM, and LMS.

The MHRM algorithm (Cai, 2008) presented earlier is a valid option for estimating a multilevel IRT model with nonlinear latent variable effects. As mentioned before, simulation studies (e.g., Cai (2010), Han and Paek (2014), and Kuo and Sheng (2016)) show that both the EM with Gauss-Hermite quadrature and the MHRM produce comparably small bias and RMSE. However, these studies also suggest that the EM with GHQ has a slightly lower bias than the MHRM implementation, especially in small datasets and as the correlation between the latent variables $\boldsymbol{\xi}$ rises. The inclusion of nonlinear latent variable effects further complicates an already complex multilevel model. Therefore, a decision was made in favor of a potentially more accurate estimation method, at the cost of a slower performance, compared to the one that the MHRM would have undoubtedly provided. Nonetheless, the MHRM algorithm should be investigated in future research.

2.3 Deduction of the Estimation Procedure

In this section, the objective function for estimating all parameters from Model (2.1) using the EM will be developed. First, the basic theoretical EM is deduced. In order to better understand the derivation and be able to implement and apply it, a different – more intuitive – perspective of the EM algorithm will be presented, which was first introduced by Neal and Hinton (1998) and revised by Dellaert (2002). They deduced the EM as an algorithm that iteratively maximizes a lower bound of the observed data likelihood.

First, the EM will be defined for a single-level model with nonlinear latent variable effects by creating a lower bound to the observed data log-likelihood. This perspective will then be compared to Dempster et al. (1977), as discussed in Section 1.6. Finally, the extension to the complete multilevel IRT model with nonlinear effects is presented.

2.3.1 The EM Objective Function

In Section 1.6, the general objective function of the EM from Dempster et al. (1977) for a model without hierarchical structure was given in (1.27) as

$$\mathcal{Q}(\boldsymbol{\omega}, \boldsymbol{\omega}') = \mathbb{E}\left[\log(f(\boldsymbol{Y}, \boldsymbol{\xi}|\boldsymbol{\omega}))|\boldsymbol{Y}, \boldsymbol{\omega}'\right].$$

The parameters in $\boldsymbol{\omega}$ are estimated conditional on the fixed parameter values in $\boldsymbol{\omega}'$. The Model (2.1) that needs to be estimated now, is a multilevel model with two kinds of latent variables ($\boldsymbol{\xi}_{jk}$ belonging to the individuals and the random intercept u_k), which can be interpreted as missing values since they cannot be observed directly. For Model (2.1) the general objective function is then given as

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\omega}, \boldsymbol{\omega}') =& \mathbb{E}\left[\log(f(\boldsymbol{Y}, \xi, u|\boldsymbol{\omega}))|\boldsymbol{Y}, \boldsymbol{\omega}'\right] \\ =& \int_u \int_{\boldsymbol{\xi}} \log(f(\boldsymbol{Y}, \xi, u|\boldsymbol{\omega}))P(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega}')P(u|\boldsymbol{Y}, \boldsymbol{\omega}')d\boldsymbol{\xi}du \\ =& \int_u P(u|\boldsymbol{Y}, \boldsymbol{\omega}') \left[\int_{\boldsymbol{\xi}} \log(f(\boldsymbol{Y}, \xi, u|\boldsymbol{\omega}))P(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega}')d\boldsymbol{\xi}\right] du.\end{aligned}$$

However, it conceals important aspects that must be understood to actually implement the estimation procedure. In the following sections, the EM is fully deduced from a different perspective, resulting in the same basic formula. However, the new procedure will be more detailed, give more information on the inner workings of the EM and provide insight into numerical aspects that need to be taken into account.

2.3.2 EM for an IRT Model with Nonlinear Latent Variable Effects

The overall goal is the estimation of Model (2.1). For better understanding, the deduction is first conducted for a model *without a random intercept* where only one cluster is present.

The Log-Likelihood

A straightforward way would be to use MML and maximize the likelihood / posterior probability $L(\tilde{\boldsymbol{\omega}}|\boldsymbol{Y})$ of the reduced parameter vector $\tilde{\boldsymbol{\omega}}$ without the variance σ^2 of the random intercept

$$\tilde{\boldsymbol{\omega}} = (\gamma_1, \dots, \gamma_I, \Omega_1^{11}, \dots, \Omega_1^{nn}, \dots, \Omega_I^{11}, \dots, \Omega_I^{nn}, \delta_1, \dots, \delta_I, \boldsymbol{\Sigma}_\xi)$$

given the data $\boldsymbol{Y}$. Mathematically, however, a likelihood function of the parameters can be written as $P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})$, where the data $\boldsymbol{Y}$ are fixed and the parameters $\tilde{\boldsymbol{\omega}}$ are free (in contrast to a probability where the parameters are fixed and the data are free).

Therefore, here it is given as the distribution of the observed variables given the parameters, which can be rewritten using the complete data likelihood. To facilitate the calculation, the logarithm of the likelihood is taken, which results in

$$\underset{\tilde{\boldsymbol{\omega}}\in\Omega}{\arg\max} \log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}) = \underset{\tilde{\boldsymbol{\omega}}\in\Omega}{\arg\max} \log \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}). \tag{2.3}$$

A Lower Bound

The integration over the ability $\boldsymbol{\xi}$ is potentially multidimensional with the same dimensionality as $\boldsymbol{\xi}$. Unfortunately, the logarithm of a sum / integral is numerically not helpful. However, using Jensen's inequality, the logarithm can be brought into the sum

$$\begin{aligned}
\log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}) &= \log \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \\
&= \log \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \frac{P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})} \\
&\geq \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \log \frac{P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})} \qquad (2.4)
\end{aligned}$$

where $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$ is an arbitrary probability distribution of $\boldsymbol{\xi}$ depending on the same parameters as the likelihood. This simultaneously creates a lower bound to the log-likelihood of the observed data. The EM now uses the best possible lower bound and maximizes it to approximate $\log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})$. In other words, it uses that density $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$ that maximizes the lower bound, which is then in turn maximized. Dempster et al. (1977) showed that this procedure converges to a maximum of $\log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})$ if the maximization of the lower bound is iterated by substituting the current estimates into the density $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$.

The Optimal Lower Bound

To achieve the optimal lower bound, the function (2.4) will be maximized over $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$. To assure that the density $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$ stays normalized, a Lagrange multiplier is added to the objective function, which results in

$$\begin{aligned} G_\lambda(f) =& \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \log P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) - \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \log f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \\ & + \lambda\left(1 - \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})\right). \end{aligned}$$

Its first partial derivative with respect to $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$ is found using the functional derivative

$$\begin{aligned} \sum_{\boldsymbol{\xi}\in\Xi} \frac{\partial G_\lambda}{\partial f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}\phi(\boldsymbol{\xi}) &= \left[\frac{d}{d\epsilon} G_\lambda(f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) + \epsilon\phi(\boldsymbol{\xi}))\right]_{\epsilon=0} \\ = \Bigg[\frac{d}{d\epsilon} & \sum_{\boldsymbol{\xi}\in\Xi} (f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) + \epsilon\phi(\boldsymbol{\xi})) \log P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \\ & - \sum_{\boldsymbol{\xi}\in\Xi} (f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) + \epsilon\phi(\boldsymbol{\xi})) \log(f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) + \epsilon\phi(\boldsymbol{\xi})) \\ & + \lambda\left(1 - \sum_{\boldsymbol{\xi}\in\Xi} (f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) + \epsilon\phi(\boldsymbol{\xi}))\right)\Bigg]_{\epsilon=0} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{\xi}\in\Xi} \phi(\boldsymbol{\xi}) \log P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) - \sum_{\boldsymbol{\xi}\in\Xi} (1 + \log(f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}))\phi(\boldsymbol{\xi}) + \sum_{\boldsymbol{\xi}\in\Xi} (-\lambda\phi(\boldsymbol{\xi}))\\
&= \sum_{\boldsymbol{\xi}\in\Xi} (\log P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) - \log(f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})) - 1 - \lambda)\phi(\boldsymbol{\xi})
\end{aligned}
$$

which results in

$$
\frac{\partial G_\lambda}{\partial f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})} = \log P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) - \log(f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})) - 1 - \lambda. \tag{2.5}
$$

Setting Equation (2.5) equal to zero[1] yields

$$
\log P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) - \log(f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})) - 1 - \lambda \overset{!}{=} 0 \tag{2.6}
$$

$$
\Leftrightarrow f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) = \frac{P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{\exp(1+\lambda)}. \tag{2.7}
$$

The normalization term still needs to be defined by setting the first partial derivative of G_λ with respect to λ equal to zero and solving for $\exp(1+\lambda)$. It trivially results in

$$
\begin{aligned}
\frac{\partial G_\lambda}{\partial \lambda} &= 1 - \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \overset{!}{=} 0\\
\Leftrightarrow 1 &= \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \overset{(2.7)}{=} \frac{1}{\exp(1+\lambda)} \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})\\
\Leftrightarrow \exp(1+\lambda) &= \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}). \qquad (2.8)
\end{aligned}
$$

Substituting (2.8) into (2.7) results in the density function

$$
f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) = \frac{P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})} = \frac{P(\boldsymbol{Y}|\boldsymbol{\xi}, \tilde{\boldsymbol{\omega}})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)}{P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})} = P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}) \tag{2.9}
$$

that achieves the optimal lower bound in (2.4). The optimal density function (2.9) can be described as the posterior distribution of the latent variable $\boldsymbol{\xi}$. Since it was earlier assumed that $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\xi)$, the parameter vector $\tilde{\boldsymbol{\omega}}$ in $P(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$ was substituted with the only relevant parameters $\boldsymbol{\Sigma}_\xi$.

[1] Setting an equation equal to a different term is written as $\overset{!}{=}$ in mathematical contexts.

The reevaluation of the inequality (2.4) for $f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) = P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}) = \frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})}$

$$\begin{aligned}\sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) \log \frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{f(\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})} &= \sum_{\boldsymbol{\xi}\in\Xi} \frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})} \log \frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{\frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})}} \\ &= \sum_{\boldsymbol{\xi}\in\Xi} \frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})} \log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}) \\ &= \log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}) \qquad (2.10)\end{aligned}$$

shows that in Jensen's Inequality, even equality to the log-likelihood of the observed parameters is achieved when the posterior likelihood of the missing data is used – with the same set of parameters $\tilde{\boldsymbol{\omega}}$ in both the posterior and the likelihood.

The optimal lower bound will be denoted by

$$B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)}) = \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log \frac{P(\boldsymbol{Y},\boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})}{P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)})} \qquad (2.11)$$

where different sets of parameters – $\tilde{\boldsymbol{\omega}}^{(p)}$ and $\tilde{\boldsymbol{\omega}}$ – are included. The vector $\tilde{\boldsymbol{\omega}}^{(p)}$ are the parameters of iteration p which are fixed in the posterior probability of the latent variables. This essentially corresponds to the *expectation step* since the current values of the missing data are estimated using their conditional expectation given the data and the current parameter estimates. The *maximization step* is given by maximizing $B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ over the unknown $\tilde{\boldsymbol{\omega}}$ which only vary in the complete likelihood function $P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}})$.

Visualization of the Lower Bound

The more intuitive and visual aspect of this approach becomes clearer by observing that the optimization to obtain parameters $\tilde{\boldsymbol{\omega}}^{(p+1)}$ is initiated using the starting values $\tilde{\boldsymbol{\omega}}^{(p)}$. This in turn results in

$$B(\tilde{\boldsymbol{\omega}}^{(p)}, \tilde{\boldsymbol{\omega}}^{(p)}) = \log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}^{(p)})$$

so that $B(\tilde{\boldsymbol{\omega}}^{(p)}, \tilde{\boldsymbol{\omega}}^{(p)})$ builds a tangent to $\log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}^{(p)})$ for the current parameter estimates. After the optimization, $B(\tilde{\boldsymbol{\omega}}^{(p+1)}, \tilde{\boldsymbol{\omega}}^{(p+1)})$ again builds a tangent to $\log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}^{(p+1)})$ in $\tilde{\boldsymbol{\omega}}^{(p+1)}$.

In other words, the current objective function builds a tangent to the observed data likelihood at the optimal values of the previous iteration, which Figure 2.1 depicts for a one-dimensional parameter vector. The optimal value $\tilde{\boldsymbol{\omega}}^{(p+1)}$ of $B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ is then always an improvement on the observed data likelihood since the slopes of both functions are the same in $\boldsymbol{\omega}^{\boldsymbol{(p)}}$. The objective function $B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p+1)})$ of the next iteration again builds a tangent, now in $P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}^{(p+1)})$. The $B(\tilde{\boldsymbol{\omega}}^{(p)}, \tilde{\boldsymbol{\omega}}^{(p)})$ always form a lower bound on the observed log-likelihood, which is steadily increased until the lower bound is close enough to the (local) maximum of $\log P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}}^{(p)})$.

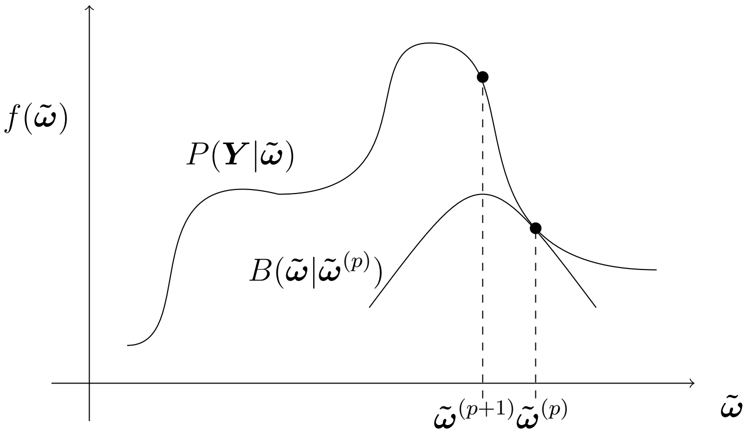
$f(\tilde{\boldsymbol{\omega}})$
$P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})$
$B(\tilde{\boldsymbol{\omega}}|\tilde{\boldsymbol{\omega}}^{(p)})$
$\tilde{\boldsymbol{\omega}}^{(p+1)}$
$\tilde{\boldsymbol{\omega}}^{(p)}$
$\tilde{\boldsymbol{\omega}}$

Figure 2.1: Illustration of the EM objective function $B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ as a tangent to the observed data likelihood $P(\boldsymbol{Y}|\tilde{\boldsymbol{\omega}})$ in the current optimal value $\tilde{\boldsymbol{\omega}}^{(p)}$.

Additionally, it can be observed that $B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ can be split

$$B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)}) = \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{Y}, \boldsymbol{\xi}|\tilde{\boldsymbol{\omega}}) - \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)})$$

into the expectancy of the complete data log-likelihood conditioned on the data and the parameters of the current iteration and into the entropy of the posterior distribution of the latent trait. Since the second term only depends on the parameters of the current iteration $\tilde{\boldsymbol{\omega}}^{(p)}$ and not on $\tilde{\boldsymbol{\omega}}$, it can be left out when maximizing $B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$, resulting in the objective function

$$B_O(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)}) = \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{Y}, \boldsymbol{\xi} | \tilde{\boldsymbol{\omega}}). \tag{2.12}$$

2.3.3 Comparison of two EM approaches

The alternative deduction of the EM in the previous section is now brought together with the perspective from Section 1.6. The examination of the EM using integrals or infinite sums as notation for the expectation facilitates handling the functions in Dempster et al. (1977).

The equivalence of the EM objective function $\mathcal{Q}(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}')$ given in Equation (1.27) and the one given in the previous section in (2.12) can easily be seen. The objective function $B_O(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ can be interpreted as the expectation of the complete data log-likelihood $\log P(\boldsymbol{Y}, \boldsymbol{\xi} | \tilde{\boldsymbol{\omega}})$ conditional on the observed data $\boldsymbol{Y}$ and the current estimates $\tilde{\boldsymbol{\omega}}^{(p)}$, so that

$$\begin{aligned} B_O(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)}) =& \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{Y}, \boldsymbol{\xi} | \tilde{\boldsymbol{\omega}}) = \mathbb{E}\left[\log(f(\boldsymbol{\xi}, \boldsymbol{Y} | \tilde{\boldsymbol{\omega}})) | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}'\right] \\ =& \mathcal{Q}(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}'). \end{aligned}$$

Consequently, the equivalence of the objective function $B_O(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ to the sum of the expectation of the posterior log-likelihood $k(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}})$, conditional on the observed data $\boldsymbol{Y}$ and on a different realization of the parameters $\tilde{\boldsymbol{\omega}}'$, and of the observed log-likelihood as in Equation (1.28) is valid here as well:

$$\begin{aligned} B(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)}) &= \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{Y}, \boldsymbol{\xi} | \tilde{\boldsymbol{\omega}}) \\ &= \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log \left(P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}) \cdot P(\boldsymbol{Y} | \tilde{\boldsymbol{\omega}})\right) \\ &= \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}) + \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi} | \boldsymbol{Y}, \tilde{\boldsymbol{\omega}}^{(p)}) \log P(\boldsymbol{Y} | \tilde{\boldsymbol{\omega}}) \\ &= H(\tilde{\boldsymbol{\omega}} | \tilde{\boldsymbol{\omega}}^{(p)}) + \log P(\boldsymbol{Y} | \tilde{\boldsymbol{\omega}}). \end{aligned} \tag{2.13}$$

Now, the still-to-be-done proof of the inequation (1.29) is much easier. For all pairs $(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}') \in (\Omega, \Omega)$

$$
\begin{aligned}
& H(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}') - H(\tilde{\boldsymbol{\omega}}', \tilde{\boldsymbol{\omega}}') \\
=& \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}') \log P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}) - \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}') \log P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}') \\
=& \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}') \log \frac{P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}})}{P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}')} \\
\overset{\text{Jensen}}{\leq} & \log \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}') \frac{P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}})}{P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}')} \\
=& \log \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \tilde{\boldsymbol{\omega}}) \\
=& 0. \qquad (2.14)
\end{aligned}
$$

The intuitive deduction of the EM as a lower bound also better explains why it is sufficient to only consider $B_O(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)}) = Q(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ even though (1.28) might suggest that $H(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ should also be included in the optimization process. Furthermore, $H(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}') \leq 0$, which justifies seeing $Q(\tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\omega}}^{(p)})$ as a lower bound in (1.28).

2.3.4 EM for a Multilevel IRT Model with Nonlinear Latent Variable effects

In Section 2.3.2, the estimation of an IRT model with nonlinear latent variable effects was deduced. However, if a hierarchical structure is present, the same deduction needs to be done twice, since not only the latent abilities but also the random intercept need to be marginalized out of the likelihood.

Furthermore, the variance of the random intercept is estimated separately from the item parameters and the variance-covariance matrix, which stabilizes the optimization. The individual objective functions are deduced and the estimation process is described.

The General Objective Function for MINoLEM

Unlike the former calculations, this time the indices for every person are included as in Model 2.1. Furthermore, the parameter vector

$$\boldsymbol{\omega} = (\gamma_1, \dots, \gamma_I, \Omega_1^{11}, \dots, \Omega_1^{nn}, \dots, \Omega_I^{11}, \dots, \Omega_I^{nn}, \delta_1, \dots, \delta_I, \boldsymbol{\Sigma}_\xi, \sigma^2)$$

needs to be extended again to include the variance of the random intercept.

The complete objective function for a multilevel IRT model with nonlinear latent variable effects can be build as in Section 2.3.2. First, the logarithm is brought into the integration over u and then into the integration over $\boldsymbol{\xi}$:

$$\begin{aligned} &\log P(\boldsymbol{Y}|\boldsymbol{\omega}) \\ =&\log \prod_{k=1}^{K} \sum_{u\in U} \prod_{j=1}^{J_k} \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega}) \\ =&\log \prod_{k=1}^{K} \sum_{u\in U} f(u|\boldsymbol{\omega}) \frac{\prod_{j=1}^{J_k} \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega})}{f(u|\boldsymbol{\omega})} \end{aligned}$$

$$\geq \sum_{k=1}^{K} \sum_{u\in U} f(u|\boldsymbol{\omega}) \left(\log \left(\prod_{j=1}^{J_k} \sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega}) \right) - \log\left(f(u|\boldsymbol{\omega})\right) \right) \tag{2.15}$$

$$\geq \sum_{k=1}^{K} \sum_{u\in U} f(u|\boldsymbol{\omega}) \left(\sum_{j=1}^{J_k} \sum_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{\xi}|\boldsymbol{\omega}) \log \frac{P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega})}{f(\boldsymbol{\xi}|\boldsymbol{\omega})} - \log\left(f(u|\boldsymbol{\omega})\right) \right) \tag{2.16}$$

For each inequality (2.15) and (2.16) it can be shown (see Section 1.6) that the lower bounds are optimal if the posterior probabilities of the random intercept $P(u|\boldsymbol{Y_k}$ and the latent variable $P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega})$ are chosen for $f(u|\boldsymbol{\omega})$

and $f(\boldsymbol{\xi}|\boldsymbol{\omega})$, respectively. The posterior probabilities are given by

$$\begin{aligned} P(u|\boldsymbol{Y_k},\boldsymbol{\omega}) &= \frac{P(\boldsymbol{Y_k},u|\boldsymbol{\omega})}{P(\boldsymbol{Y_k}|\boldsymbol{\omega})} = \frac{P(\boldsymbol{Y_k}|u,\boldsymbol{\omega})P(u|\sigma^2)}{P(\boldsymbol{Y_k}|\boldsymbol{\omega})} \\ &= \frac{\prod\limits_{j=1}^{J_k}\sum\limits_{\boldsymbol{\xi}\in\Xi}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi},u,\boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)P(u|\sigma^2)}{\sum\limits_{u\in U}\prod\limits_{j=1}^{J_k}\sum\limits_{\boldsymbol{\xi}\in\Xi}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi},u,\boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)P(u|\sigma^2)} \end{aligned}$$

and

$$\begin{aligned} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}},\boldsymbol{\omega}) &= \frac{P(\boldsymbol{Y_{jk}},\boldsymbol{\xi}|\boldsymbol{\omega})}{P(\boldsymbol{Y_{jk}}|\boldsymbol{\omega})} = \frac{P(\boldsymbol{Y_{jk}}|\boldsymbol{\xi},\boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)}{\sum_{\boldsymbol{\xi}\in\Xi}P(\boldsymbol{Y_{jk}}|\boldsymbol{\xi},\boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)} \\ &= \frac{\sum\limits_{u\in U}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi},u,\boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)P(u|\sigma^2)}{\sum\limits_{\boldsymbol{\xi}\in\Xi}\sum\limits_{u\in U}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi},u,\boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}_\xi)P(u|\sigma^2)} \end{aligned}$$

Again, since $u \sim \mathcal{N}(0,\sigma^2)$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_\xi)$, the parameter vector $\boldsymbol{\omega}$ is substituted by σ^2 and $\boldsymbol{\Sigma}_\xi$ in the distributions of the latent variables. For better readability, the variance-covariance matrix of $\boldsymbol{\xi}$ will in the future only be written as $\boldsymbol{\Sigma}$. Substituting $f(u|\boldsymbol{\omega})$ and $f(\boldsymbol{\xi}|\boldsymbol{\omega})$ in the objective function (2.16) results in

$$\begin{aligned} &\log P(\boldsymbol{Y}|\boldsymbol{\omega}) \\ =&\sum_{k=1}^{K}\sum_{u\in U}P(u|\boldsymbol{Y_k},\boldsymbol{\omega})\left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi}P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}},\boldsymbol{\omega})\log(P(\boldsymbol{Y_{jk}},\boldsymbol{\xi},u|\boldsymbol{\omega}))\right) \qquad (2.17) \\ &-\sum_{k=1}^{K}\sum_{u\in U}P(u|\boldsymbol{Y_k},\boldsymbol{\omega})\left(\log(P(u|\boldsymbol{Y_k},\boldsymbol{\omega}))+\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi}P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}},\boldsymbol{\omega})\log(P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}},\boldsymbol{\omega}))\right). \end{aligned}$$

Again, the first term in (2.17) is the expectation of the complete data likelihood conditional on the posterior likelihoods of the random intercept u and the latent ability $\boldsymbol{\xi}$. The second term is the entropy of the posterior of u (summed over all clusters) and the expected entropy of the posterior of $\boldsymbol{\xi}$ conditional on the random intercept u (and summed over all clusters).

As established for Equation (2.13), the parameter values in the posterior probabilities will be fixed to the estimates of the current iteration, which gives the objective function

$$\begin{aligned}
&B_{OF}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) \qquad (2.18)\\
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}) \left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}) \log(P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega}))\right)
\end{aligned}$$

where the second term from (2.17) can be left out, since it only contains the fixed parameters $\boldsymbol{\omega}^{(p)}$ and is therefore constant within an optimization. This objective function can also be described as the expectation of the complete data log-likelihood conditional on the posteriors of the latent ability and the random intercept.

Individual Objective Functions for Each Parameter Set

Raudenbush and Bryk (2002) describe that the estimation of the variance components depends on the information about the fixed parameters and vice versa. That is why they build two separate likelihoods for the variance and fixed effects and iterate their estimation until convergence is achieved.

However, this only applies to the variance of the random intercept. Simulations have shown that the estimation performs significantly better if the variance-covariance matrix $\boldsymbol{\Sigma}$ of the latent variable $\boldsymbol{\xi}$ is estimated together with the item parameters.

Accordingly, the estimation of the variance of the random intercept and of the fixed parameters together with $\boldsymbol{\Sigma}$ from Model (2.1) should be separated and iterated as well. To better show the estimation of the variance-covariance matrix $\boldsymbol{\Sigma}$ and of the variance σ^2, the parameter vector $\boldsymbol{\omega}$ will, once again, be redefined without them as

$$\boldsymbol{\omega} = (\gamma_1, \ldots, \gamma_I, \Omega_1^{11}, \ldots, \Omega_1^{nn}, \ldots, \Omega_I^{11}, \ldots, \Omega_I^{nn}, \delta_1, \ldots, \delta_I)$$

The objective function (2.18) will be defined once with fixed variance $\sigma^{(p)}$ and once with fixed parameters $\boldsymbol{\omega}^{(p)}$ and $\boldsymbol{\Sigma}^{(p)}$. The objective function to estimate $\boldsymbol{\omega}$ and $\boldsymbol{\Sigma}$ is given by

$$\begin{aligned}
&B_{OF-\boldsymbol{\omega}}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) \\
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega}))\right) \\
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \left[\left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(P(\boldsymbol{Y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega}))\right)\right. \\
&\quad + \left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(P(\boldsymbol{\xi}|\boldsymbol{\Sigma}))\right) \qquad (2.19) \\
&\quad \left. + \left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(P(u|\sigma^{(p)}))\right)\right].
\end{aligned}$$

The probability of $\boldsymbol{Y_{jk}}$ is transformed to be dependent on the latent variables $\boldsymbol{\xi}$ and u by multiplying with their respective probabilities. The last term does not depend on $\boldsymbol{\omega}$ or $\boldsymbol{\Sigma}$ and can be ignored in the optimization. For the objective function to estimate σ, the same considerations lead to

$$\begin{aligned}
&B_{OF-\sigma}(\sigma, \sigma^{(p)}) \\
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \\
&\quad \left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(P(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u|\boldsymbol{\omega}^{(p)}))\right) \\
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \\
&\quad \left[\left(\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(P(\boldsymbol{Y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)}))\right)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \cdot \log(P(u|\sigma)) \sum_{j=1}^{J_k} 1, \qquad (2.20)
\end{aligned}$$

where a property of densities was applied:

$$\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}},\boldsymbol{\omega}^{(p)},\boldsymbol{\Sigma}^{(p)})\log(P(u|\sigma)) = \log(P(u|\sigma))\sum_{j=1}^{J_k} 1.$$

It is important to note that additive last term of the objective function is the product of the posterior probability of the random intercept and the logarithm of its assumed distribution. This is what primarily drives the optimization of the variance between the clusters. The same is true for the product of the posterior probability of the random variable $\boldsymbol{\xi}$ and the logarithm of its assumed distribution in Equation (2.19): This term is mainly responsible for the estimation of the variance-covariance matrix $\boldsymbol{\Sigma}$.

Ultimately, the EM for a multilevel IRT model with nonlinear latent variable effects consists of 3 steps:

1. Expectation Step: calculate the posterior probabilities $P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}},\boldsymbol{\omega}^{(p)},\boldsymbol{\Sigma}^{(p)})$ and $P(u|\boldsymbol{Y_k},\boldsymbol{\omega}^{(p)},\sigma^{(p)})$
2. Maximization Step I: maximize the likelihood $B_{OF-\boldsymbol{\omega}}(\boldsymbol{\omega},\boldsymbol{\omega}^{(p)})$ with fixed random intercept $\sigma^{(p)}$
3. Maximization Step II: maximize the likelihood $B_{OF-\sigma}(\sigma,\sigma^{(p)})$ with fixed item parameters $\boldsymbol{\omega}^{(p+1)}$ and fixed variance-covariance matrix $\boldsymbol{\Sigma}^{(p+1)}$ from step 2.

These are iterated until convergence is reached.

2.4 Numerical Application of the Estimation Procedure

In the previous sections, the theoretical EM algorithm for a multilevel IRT model with nonlinear latent variable effects was introduced. In order to apply the procedure, a few additional steps must be taken into account. The integrals need to be calculated numerically by applying GHQ. The derivatives of the objective functions are needed to optimize them more efficiently. At the beginning of the estimation, starting values need to be chosen, which will be discussed here. The convergence of the implementation will be examined as well.

2.4.1 Numerical Integration

The chosen method of numerical integration is GHQ. A possible alternative is adaptive GHQ (Liu & Pierce, 1994), which achieves reasonable accuracy with significantly fewer quadrature points. However, since this estimation procedure is new, computationally complex, and GHQ theoretically achieves arbitrary accuracy, a decision was made in favor of the potentially more accurate GHQ.

The GHQ can be applied to integrals that have the form

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\, dx \approx \sum_{i=1}^{n} \alpha_i f(x_i)$$

where α_i are the weights, x_i are the quadrature points, and n is the number of quadrature points chosen by the user. The more quadrature points are chosen, the more accurate the integral is estimated, but the more computationally demanding the problem becomes. The quadrature points are gained by finding the roots of the physicists' Hermite polynomial $H_n(x)$ of the order n. The corresponding weights are given by

$$\alpha_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_i)]^2}.$$

To apply the GHQ, a Cholesky Decomposition must be used, as the GHQ is built for standard normally distributed variables, but the (2-dimensional) latent variable and the random intercept are assumed to follow a normal distribution with

$$\boldsymbol{\xi} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_{11}^2 & r\\ r & \sigma_{22}^2\end{pmatrix}\right) \quad \text{and} \quad u \sim \mathcal{N}(0, \sigma^2).$$

However, a variance-covariance matrix can be written as

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$$

where $\boldsymbol{L}$ is a lower triangular matrix. This representation is unique and possible since $\boldsymbol{\Sigma}$ – as a variance-covariance matrix – is positive definite. In the case of a 2×2 variance-covariance matrix, $\boldsymbol{L}$ can be deduced by

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' = \begin{pmatrix}\sigma_{11}^2 & r\\ r & \sigma_{22}^2\end{pmatrix} = \begin{pmatrix}a & 0\\ b & c\end{pmatrix}\begin{pmatrix}a & b\\ 0 & c\end{pmatrix} = \begin{pmatrix}a^2 & ab\\ ab & b^2 + c^2\end{pmatrix}$$

$$\Rightarrow \begin{cases}\sigma_{11} = a\\ r = a \cdot b\\ \sigma_{22}^2 = b^2 + c^2\end{cases} \quad \Leftrightarrow \begin{cases}a = \sigma_{11}\\ b = \frac{r}{\sigma_{11}}\\ c = \sqrt{\sigma_{22}^2 - \left(\frac{r}{\sigma_{11}}\right)^2}\end{cases}$$

which results in

$$\boldsymbol{L} = \begin{pmatrix}\sigma_{11} & 0\\ \frac{r}{\sigma_{11}} & \sqrt{\sigma_{22}^2 - \left(\frac{r}{\sigma_{11}}\right)^2}\end{pmatrix}$$

Furthermore, a multivariate standard normally distributed vector $\boldsymbol{Z} = (z_1, \ldots, z_n)'$ can be transformed so that $\boldsymbol{L} \cdot \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{L}\boldsymbol{L}') = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Using the Cholesky transformation, the GHQ can be applied to integrals that include an arbitrary normal distribution since the GHQ needs integrals of the form

$$\int e^{-x^2} f(x) dx \tag{2.21}$$

and approximates the integral with a sum

$$\sum_{i=1}^{n} w_i f(x_i),$$

where the weights w_i and the quadrature points x_i are both determined by Hermite polynomials.

The application of GHQ is first illustrated with the objective function (2.19) of the item parameters $\boldsymbol{\omega}$ and of the variance-covariance matrix $\boldsymbol{\Sigma}$ of the latent variables, presented here in a more detailed form:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)})\cdot \\
&\quad\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log[L(\boldsymbol{Y_{jk}}, \boldsymbol{\xi}, u, |\boldsymbol{\omega})]\right]
\end{aligned}\tag{2.22}$$

$$\begin{aligned}
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)})\cdot \\
&\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} \frac{\sum\limits_{u\in U}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}^{(p)})P(u|\sigma^{(p)})}{\sum\limits_{\boldsymbol{\xi}\in\Xi}\sum\limits_{u\in U}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)})P(\boldsymbol{\xi}|\boldsymbol{\Sigma}^{(p)})P(u|\sigma^{(p)})}\cdot\right. \\
&\quad\left.\log\Big[L(\boldsymbol{Y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega})P(\boldsymbol{\xi}|\boldsymbol{\Sigma})P(u|\sigma^{(p)})\Big]\right]
\end{aligned}$$

$$\begin{aligned}
=&\sum_{k=1}^{K}\sum_{u\in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)})\cdot \\
&\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi} \frac{\displaystyle\sum_{u\in U}\prod_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)})\frac{exp(-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\Sigma}_{(p)}^{-1}\boldsymbol{\xi})}{\sqrt{4\pi^2\det(\boldsymbol{\Sigma_{(p)}})}}\frac{exp(-\frac{u^2}{2\sigma_{(p)}^2})}{\sqrt{2\pi\sigma_{(p)}^2}}}{\displaystyle\sum_{\boldsymbol{\xi}\in\Xi}\sum_{u\in U}\prod_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)})\frac{exp(-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\Sigma}_{(p)}^{-1}\boldsymbol{\xi})}{\sqrt{4\pi^2\det(\boldsymbol{\Sigma_{(p)}})}}\frac{exp(-\frac{u^2}{2\sigma_{(p)}^2})}{\sqrt{2\pi\sigma_{(p)}^2}}}\cdot\right.
\end{aligned}\tag{2.23}$$

$$\left.\log\left[L(\boldsymbol{Y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega})\frac{exp(-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi})}{\sqrt{4\pi^2\det(\boldsymbol{\Sigma})}}\frac{exp(-\frac{u^2}{2\sigma_{(p)}^2})}{\sqrt{2\pi\sigma_{(p)}^2}}\right]\right],\tag{2.24}$$

where the fractions of the distributions in (2.23) can be reduced, since the terms $\sqrt{4\pi^2 \det(\boldsymbol{\Sigma_{(p)}})}$ and $\sqrt{2\pi\sigma^2_{(p)}}$ occur in the nominator and the denominator.

Unfortunately, the objective function is not in a form as in Equation (2.21), yet. First, the random effects need to be reparameterized to become standard normally distributed variables. As established before, a multivariate standard normally distributed n-dimensional vector $\boldsymbol{Z}$ can be written as $\boldsymbol{L} \cdot \boldsymbol{Z} = \boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$. The random intercept can be transformed by $\sigma \cdot z = u \sim \mathcal{N}(0, \sigma^2)$.

The Cholesky transformation is applied by integrating by substitution. In the case of a one-dimensional variable

$$\int_{\varphi(a)}^{\varphi(b)} f(u)\, du = \int_a^b f(\varphi(x))\varphi'(x)\, dx \tag{2.25}$$

holds with $\varphi(x) = u$. In the multivariate case

$$\int_{\varphi(U)} f(\boldsymbol{v})\, d\boldsymbol{v} = \int_U f(\varphi(\boldsymbol{u}))\, |\det \varphi'(\boldsymbol{u})|\; d\boldsymbol{u} \tag{2.26}$$

holds with $\varphi(\boldsymbol{u})$ being the multivariate transformation function. However, the reparameterising functions do not only contain the Cholesky Decomposition but also the factor $\sqrt{2}$ to obtain the function $\exp(-x^2)$ as in Equation (2.21).

The needed derivatives from $\varphi_\xi(\boldsymbol{\xi}) = \sqrt{2}\boldsymbol{L}_{(p)} \cdot \boldsymbol{Z}$ and $\varphi_u(u) = \sqrt{2}\sigma^{(p)} \cdot z$ are given by $\varphi'_\xi(\boldsymbol{\xi}) = \sqrt{2}\boldsymbol{L}_{(p)}$ and $\varphi'_u(u) = \sqrt{2}\sigma^{(p)}$, respectively. According to Equations (2.25) and (2.26), the substitution of the latent variables will result in a multiplication of the integrand with

$$\sqrt{2}|\det(\boldsymbol{L}_{(p)})| \cdot \sqrt{2}\sigma^{(p)} = 2|\det(\boldsymbol{L}_{(p)})| \cdot \sigma^{(p)}. \tag{2.27}$$

It needs to be noted that variance-covariance matrix is present in the likelihood as fixed values $\boldsymbol{\Sigma}^{(p)}$ and as parameters $\boldsymbol{\Sigma}$, while the variance of the random intercept $\sigma^{(p)}$ occurs only fixed, since it is not estimated with this likelihood. Therefore, the substitutions need to be done with the fixed parameter values of the previous iteration so that the posterior probabilities still only depend on the fixed estimates, for which $\boldsymbol{\Sigma}^{(p)} = \boldsymbol{L}_{(p)}\boldsymbol{L}'_{(p)}$ needs to be defined.

Now, the Cholesky Decomposition can be applied, and the GHQ can be presented for Equation (2.22). The nominator and denominator in Equation (2.23) differ only in the integration over $\boldsymbol{\xi}$. Therefore, only the nominator is considered at first to simplify the portrayal of the equations. The first step is the substitution of the latent variables (omitting the multiplication with (2.27)), followed by applying $\boldsymbol{\Sigma}^{(p)} = \boldsymbol{L}_{(p)}\boldsymbol{L}'_{(p)}$:

$$\begin{aligned}
&\sum_{\boldsymbol{\xi}\in\Xi}\sum_{u\in U}\prod_{i=1}^{I} P(Y_{ijk}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)}) exp(-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\Sigma}^{-1}_{(p)}\boldsymbol{\xi}) exp(-\frac{u^2}{2\sigma^2_{(p)}}) \\
=&\sum_{\boldsymbol{Z}\in\mathcal{Z}_\xi}\sum_{z\in\mathcal{Z}_u}\prod_{i=1}^{I} P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)})\cdot \qquad (2.28)\\
&\quad exp(-\boldsymbol{Z}'\boldsymbol{L}'_{(p)}\boldsymbol{\Sigma}^{-1}_{(p)}\boldsymbol{L}_{(p)}\boldsymbol{Z}) exp(-\frac{\sigma^2_{(p)}z^2}{\sigma^2_{(p)}}) \\
=&\sum_{\boldsymbol{Z}\in\mathcal{Z}_\xi}\sum_{z\in\mathcal{Z}_u}\prod_{i=1}^{I} P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)}) exp(-\boldsymbol{Z}'\boldsymbol{Z}) exp(-z^2) \qquad (2.29)
\end{aligned}$$

Next, the term in (2.24) is considered:

$$\begin{aligned}
&\log\left[L(\boldsymbol{Y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega})\frac{exp(-\frac{1}{2}\boldsymbol{\xi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi})}{\sqrt{4\pi^2\det(\boldsymbol{\Sigma})}}\frac{exp(-\frac{u^2}{2\sigma^2_{(p)}})}{\sqrt{2\pi\sigma^2_{(p)}}}\right] \\
&\log\left[L(\boldsymbol{Y_{jk}}|\sqrt{2}\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega})\frac{exp(-\boldsymbol{Z}'\boldsymbol{L}'_{(p)}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}_{(p)}\boldsymbol{Z})}{\sqrt{4\pi^2\det(\boldsymbol{\Sigma})}}\cdot\frac{exp(-\frac{\sigma^2_{(p)}z^2}{\sigma^2_{(p)}})}{\sqrt{2\pi\sigma^2_{(p)}}}\right] \\
&\log(L(\boldsymbol{Y_{jk}}|\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega})) - \boldsymbol{Z}'\boldsymbol{L}'_{(p)}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}_{(p)}\boldsymbol{Z} - \log(\sqrt{4\pi^2\det(\boldsymbol{\Sigma})}) \\
&\quad - z^2 - \log(\sqrt{2\pi\sigma^2_{(p)}}) \qquad (2.30)
\end{aligned}$$

The terms -z^2-log($\sqrt{2\pi\sigma^2_{(p)}}$) will only be multiplied by the posterior probabilities. They will not depend on $\boldsymbol{\omega}$ or $\boldsymbol{\Sigma}$ and can be left out in the optimization. Combining the calculations in Equations (2.29), (2.30), and (2.27) results in the complete objective function for (2.19):

$$
\begin{aligned}
=&\sum_{k=1}^{K}\sum_{z\in\mathcal{Z}_u} P(z|\boldsymbol{Y_k},\boldsymbol{\omega}^{(p)},\sigma^{(p)})\cdot \\
&\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{Z}\in\mathcal{Z}_\xi}\frac{\sum\limits_{z\in\mathcal{Z}_u}\prod\limits_{i=1}^{I}P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}\boldsymbol{Z},\sqrt{2}\sigma_{(p)}z,\boldsymbol{\omega}^{(p)})exp(-\boldsymbol{Z}'\boldsymbol{Z})exp(-z^2)}{\sum\limits_{\boldsymbol{Z}\in\mathcal{Z}_\xi}\sum\limits_{z\in\mathcal{Z}_u}\prod\limits_{i=1}^{I}P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}\boldsymbol{Z},\sqrt{2}\sigma_{(p)}z,\boldsymbol{\omega}^{(p)})exp(-\boldsymbol{Z}'\boldsymbol{Z})exp(-z^2)}\right. \\
&\quad\Big[\log(L(\boldsymbol{Y_{jk}}|\boldsymbol{L}_{(p)}\boldsymbol{Z},\sqrt{2}\sigma_{(p)}z,\boldsymbol{\omega}))-\boldsymbol{Z}'\boldsymbol{L}_{(p)}'\boldsymbol{\Sigma}^{-1}\boldsymbol{L}_{(p)}\boldsymbol{Z} \\
&\quad\left.-\log(\sqrt{4\pi^2\det(\boldsymbol{\Sigma})})\Big]\right]2|\det(\boldsymbol{L}_{(p)})|\cdot\sigma^{(p)} \qquad (2.31)
\end{aligned}
$$

Since this function will be optimized, the constant term $2\cdot|\det(\boldsymbol{L}_{(p)})|\cdot\sigma^{(p)}$ can be left out. Furthermore, the term $exp(-\boldsymbol{Z}'\boldsymbol{Z})$ is necessary to apply GHQ to the integration over $\boldsymbol{Z}$ (formerly $\boldsymbol{\xi}$). The term for applying GHQ to the integration over z (formerly u) results from making the same conversions for the posterior $P(z|\boldsymbol{Y_k},\boldsymbol{\omega}^{(p)},\sigma^{(p)})$. Now the objective function has a form in which GHQ can be applied. Every integral has the form $\int f(x)e^{-x^2}$ and can be approximated by introducing quadrature points (for a two-dimensional latent variable $\boldsymbol{\xi}$) $Q_\xi^{(p_1)}$ $(p_1=1,\dots,P_1)$ for $\boldsymbol{Z_1}$, $Q_\xi^{(p_2)}$ $(p_2=1,\dots,P_2)$ for $\boldsymbol{Z_2}$, $Q_\sigma^{(p_3)}$ $(p_3=1,\dots,P_3)$ for z, and corresponding weights $\alpha_\xi^{(p_1)}$, $\alpha_\xi^{(p_2)}$, and $\alpha_\sigma^{(p_3)}$. For a two-dimensional latent variable, this results in the approximation

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\frac{\prod\limits_{j=1}^{K}\sum\limits_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum\limits_{p_2}^{P_2}\alpha_\xi^{(p_2)}\prod\limits_{i=1}^{I}P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)},Q_\xi^{(p_2)})',\sigma_{(p)}Q_\sigma^{(p_3)},\boldsymbol{\omega}^{(p)})}{\sum\limits_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\prod\limits_{j=1}^{K}\sum\limits_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum\limits_{p_2}^{P_2}\alpha_\xi^{(p_2)}\prod\limits_{i=1}^{I}P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)},Q_\xi^{(p_2)})',\sigma_{(p)}Q_\sigma^{(p_3)},\boldsymbol{\omega}^{(p)})}\cdot \\
&\left[\sum_{j=1}^{J_k}\sum_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum_{p_2}^{P_2}\alpha_\xi^{(p_2)}\frac{\sum\limits_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\prod\limits_{i=1}^{I}P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)},Q_\xi^{(p_2)})',\sigma_{(p)}Q_\sigma^{(p_3)},\boldsymbol{\omega}^{(p)})}{\sum\limits_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum\limits_{p_2}^{P_2}\alpha_\xi^{(p_2)}\sum\limits_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\prod\limits_{i=1}^{I}P(Y_{ijk}|\sqrt{2}\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)},Q_\xi^{(p_2)})',\sigma_{(p)}Q_\sigma^{(p_3)},\boldsymbol{\omega}^{(p)})}\right. \\
&\Bigg[\log\Big(L(\boldsymbol{Y_{jk}}|\sqrt{2}\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)},Q_\xi^{(p_2)})',\sigma_{(p)}Q_\sigma^{(p_3)},\boldsymbol{\omega},\boldsymbol{\Sigma},\sigma^{(p)})\Big) \\
&\left.-\Big((Q_\xi^{(p_1)},Q_\xi^{(p_2)})\boldsymbol{L}_{(p)}'\boldsymbol{\Sigma}^{-1}\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)},Q_\xi^{(p_2)})')+\log(\sqrt{4\pi^2\det(\boldsymbol{\Sigma})})\Big)\Bigg]\right].
\end{aligned}
$$

For the objective function of the variance σ of the random intercept u in Equation (2.20) the Cholesky Decomposition is given by

$$
\begin{aligned}
&B_{OF-\sigma}(\sigma, \sigma^{(p)}) \\
=&\sum_{k=1}^{K}\sum_{u \in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \cdot \left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \log(L(\boldsymbol{Y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega}^{(p)})\right] \\
&+\sum_{k=1}^{K}\sum_{u \in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \cdot \log(P(u|\sigma)) \\
=&\sum_{k=1}^{K}\sum_{z \in \mathcal{Z}_u} P(z|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) \cdot \\
&\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{Z} \in \mathcal{Z}_\xi} \frac{\sum\limits_{z \in \mathcal{Z}_u}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)}) exp(-\boldsymbol{Z}'\boldsymbol{Z}) exp(-z^2)}{\sum\limits_{\boldsymbol{Z} \in \mathcal{Z}_\xi}\sum\limits_{z \in \mathcal{Z}_u}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)}) exp(-\boldsymbol{Z}'\boldsymbol{Z}) exp(-z^2)}\right. \\
&\left.\left(\log\left(L(\boldsymbol{Y_{jk}}|\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)})\right) - \frac{\sigma^2_{(p)}}{\sigma^2}z^2 - \log(\sqrt{2\pi\sigma^2})\right)\right]
\end{aligned}
$$

The application of GHQ (for a two-dimensional latent variable $\boldsymbol{\xi}$) with quadrature points $Q_\xi^{(p_1)}$ ($p_1 = 1, \ldots, P_1$) for $\boldsymbol{Z_1}$, $Q_\xi^{(p_2)}$ ($p_2 = 1, \ldots, P_2$) for $\boldsymbol{Z_2}$, $Q_\sigma^{(p_3)}$ ($p_3 = 1, \ldots, P_3$) for z, and corresponding weights $\alpha_\xi^{(p_1)}$, $\alpha_\xi^{(p_2)}$, and $\alpha_\sigma^{(p_3)}$ results in

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{p_3}^{P_3} \alpha_\sigma^{(p_3)} \frac{\prod\limits_{j=1}^{K}\sum\limits_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum\limits_{p_2}^{P_2}\alpha_\xi^{(p_2)}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)}, Q_\xi^{(p_2)})', \sigma_{(p)}Q_\sigma^{(p_3)}, \boldsymbol{\omega}^{(p)})}{\sum\limits_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\prod\limits_{j=1}^{K}\sum\limits_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum\limits_{p_2}^{P_2}\alpha_\xi^{(p_2)}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)}, Q_\xi^{(p_2)})', \sigma_{(p)}Q_\sigma^{(p_3)}, \boldsymbol{\omega}^{(p)})} \cdot \\
&\left[\sum_{j=1}^{J_k}\sum_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum_{p_2}^{P_2}\alpha_\xi^{(p_2)} \frac{\sum\limits_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)}, Q_\xi^{(p_2)})', \sigma_{(p)}Q_\sigma^{(p_3)}, \boldsymbol{\omega}^{(p)})}{\sum\limits_{p_1}^{P_1}\alpha_\xi^{(p_1)}\sum\limits_{p_2}^{P_2}\alpha_\xi^{(p_2)}\sum\limits_{p_3}^{P_3}\alpha_\sigma^{(p_3)}\prod\limits_{i=1}^{I} P(Y_{ijk}|\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)}, Q_\xi^{(p_2)})', \sigma_{(p)}Q_\sigma^{(p_3)}, \boldsymbol{\omega}^{(p)})}\right. \\
&\left.\left(\log\left(L(\boldsymbol{Y_{jk}}|\boldsymbol{L}_{(p)}(Q_\xi^{(p_1)}, Q_\xi^{(p_2)})', \sigma_{(p)}Q_\sigma^{(p_3)}, \boldsymbol{\omega}^{(p)})\right) - \left(\frac{\sigma^2_{(p)}}{\sigma^2}(Q_\sigma^{(p_3)})^2 + \log(\sqrt{2\pi\sigma^2})\right)\right)\right].
\end{aligned}
$$

However, there is a very important detail that needs to be changed in the implementation. From the given deduction, it follows that the log-likelihood of the data $\log(L(\boldsymbol{Y_{jk}}|\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)}))$ does not contain the parameter

σ (only $\sigma_{(p)}$) that is supposed to be estimated. The EM, however, is built so that the parameters to be estimated are set freely in the complete log-likelihood and then the expectation is taken over the posterior probabilities of the latent variables with all parameters fixed. Therefore, the *fixed* $\sigma_{(p)}$ in $\log(L(\boldsymbol{Y_{jk}}|\boldsymbol{L}_{(p)}\boldsymbol{Z}, \sqrt{2}\sigma_{(p)}z, \boldsymbol{\omega}^{(p)}))$ has to be *set to* σ instead, but not in $\log(P(u|\sigma)) = \frac{\sigma^2_{(p)}}{\sigma^2}z^2 + \log(\sqrt{2\pi\sigma^2})$ where σ is already present.

The optimal number of quadrature points for the latent variables highly depends on factors such as the complexity of the model and the size of the dataset. Therefore, it will be increased with rising numbers of iterations, which improves the estimation as noted by Stanley (2017).

2.4.2 Derivatives

The derivatives of the objective function are needed for optimization with quasi-Newton algorithms. The derivatives are built with the following observations. First, using local independence the log-likelihood is given by

$$\begin{aligned}\log(L(\boldsymbol{y_{jk}}|\boldsymbol{\xi}, u, \boldsymbol{\omega})) = \sum_{i=1}^{I} \log(L(y_{ijk}|\boldsymbol{\xi}, u, \boldsymbol{\omega}_i))\\ = \sum_{i=1}^{I} y_{ijk} \log(L(y_{ijk} = 1|\boldsymbol{\xi}, u, \boldsymbol{\omega}_i))\\ + \sum_{i=1}^{I} (1 - y_{ijk}) \log(1 - L(y_{ijk} = 1|\boldsymbol{\xi}, u, \boldsymbol{\omega}_i)). \qquad (2.32)\end{aligned}$$

Since the parameters (that are not fixed to the previous estimates) only occur within the likelihood of the data given the latent variables, the derivatives of $B_{OF-\omega}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)})$ can be built based on the derivatives of (2.32). Second, the converse probability of a logistic function merely changes the sign of the exponent

$$1 - \frac{1}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

and furthermore, the derivations of the logarithm of a logistic function is again a logistic function

$$\begin{aligned}\frac{\partial}{\partial x}\ln\left(\frac{1}{1+\exp(-f(x))}\right) &= \frac{\frac{\partial}{\partial x}f(x)}{1+\exp(f(x))}\\ \frac{\partial}{\partial x}\ln\left(\frac{1}{1+\exp(f(x))}\right) &= \frac{-\frac{\partial}{\partial x}f(x)}{1+\exp(-f(x))}.\end{aligned}$$

Using these considerations, the derivatives of $B_{OF-\omega}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)})$ after the item parameters in $\boldsymbol{\omega}$ can be given. Let $\gamma_{i,r}$ be the coefficient of the r-th latent variable of the n-dimensional vector $\boldsymbol{\xi}$ and let $\Omega_{i,(n_1,n_2)}$ be the coefficient of the interaction / quadratic effect of the n_1-th latent variable with the n_2-th latent variable. Then:

$$\begin{aligned}&\frac{\partial}{\partial \delta_i}B_{OF-\omega}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)})\\ =&\sum_{k=1}^{K}\sum_{u\in U}P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)})\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi}P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)})\cdot\right.\\ &\left.\left(\frac{-y_{ijk}}{1+\exp(\boldsymbol{\xi}'\boldsymbol{\gamma}_i+\boldsymbol{\xi}\boldsymbol{\Omega}_i\boldsymbol{\xi}'-\delta_i+u)}+\frac{1-y_{ijk}}{1+\exp(-(\boldsymbol{\xi}'\boldsymbol{\gamma}_i+\boldsymbol{\xi}\boldsymbol{\Omega}_i\boldsymbol{\xi}'-\delta_i+u))}\right)\right],\end{aligned}$$

$$\begin{aligned}&\frac{\partial}{\partial \gamma_{i,r}}B_{OF-\omega}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) \qquad (2.33)\\ =&\sum_{k=1}^{K}\sum_{u\in U}P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)})\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi}P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)})\cdot\right.\\ &\left.\left(\frac{y_{ijk}\cdot\xi_n}{1+\exp(\boldsymbol{\xi}'\boldsymbol{\gamma}_i+\boldsymbol{\xi}\boldsymbol{\Omega}_i\boldsymbol{\xi}'-\delta_i+u)}-\frac{(1-y_{ijk})\cdot\xi_n}{1+\exp(-(\boldsymbol{\xi}'\boldsymbol{\gamma}_i+\boldsymbol{\xi}\boldsymbol{\Omega}_i\boldsymbol{\xi}'-\delta_i+u))}\right)\right],\end{aligned}$$

and in the case of a two-dimensional latent variable with interaction present

$$\begin{aligned}&\frac{\partial}{\partial \Omega_{i,(n_1,n_2)}}B_{OF-\omega}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) \qquad (2.34)\\ =&\sum_{k=1}^{K}\sum_{u\in U}P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)})\left[\sum_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi}P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)})\cdot\right.\\ &\left.\left(\frac{y_{ijk}\cdot\xi_{n_1}\xi_{n_2}}{1+\exp(\boldsymbol{\xi}'\boldsymbol{\gamma}_i+\boldsymbol{\xi}\boldsymbol{\Omega}_i\boldsymbol{\xi}'-\delta_i+u)}-\frac{(1-y_{ijk})\cdot\xi_{n_1}\xi_{n_2}}{1+\exp(-(\boldsymbol{\xi}'\boldsymbol{\gamma}_i+\boldsymbol{\xi}\boldsymbol{\Omega}_i\boldsymbol{\xi}'-\delta_i+u))}\right)\right].\end{aligned}$$

Finally, the derivative of the objective function (2.31) with respect to the correlation ρ – if estimated – can be given. The correlation – as a parameter to be estimated – is only present in the variance-covariance matrix $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{L}_{(p)}$ includes the correlation only as a fixed value since the substitution (using the Cholesky Decomposition) is done with the fixed values. Therefore, only the term $(\boldsymbol{Z}'\boldsymbol{L}'_{(p)}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}_{(p)}\boldsymbol{Z}) + \log(\sqrt{4\pi^2 \det(\boldsymbol{\Sigma})})$ in the objective function (2.31) needs to be derived, which will now be done in the case of a two-dimensional standard normally distributed latent variable $\boldsymbol{Z} = (Z_1, Z_2)$. Furthermore, the variance-covariance matrix will be set to $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The variances in $\boldsymbol{\Sigma}$ are set to 1 to identify the model.

$$\begin{aligned}
&\frac{\partial}{\partial\rho}\left[\boldsymbol{Z}'\boldsymbol{L}'_{(p)}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}_{(p)}\boldsymbol{Z} + \log\left(\sqrt{4\pi^2\det(\boldsymbol{\Sigma})}\right)\right] \\
=&\frac{\partial}{\partial\rho}\left[\begin{pmatrix} Z_1 & Z_2 \end{pmatrix}\begin{pmatrix} 1 & \rho_{(p)} \\ 0 & \sqrt{1-\rho_{(p)}^2} \end{pmatrix}\begin{pmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \rho_{(p)} & \sqrt{1-\rho_{(p)}^2} \end{pmatrix}\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}\right. \\
&\qquad \left. + 0.5\log\left(4\pi^2\det\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)\right] \\
=&\frac{\partial}{\partial\rho}\left[\frac{1}{(\rho^2-1)}\Big[(-\rho_{(p)}^2 + 2\rho_{(p)}\rho - 1)Z_1^2 + (2\sqrt{1-\rho_{(p)}^2}(\rho-\rho_{(p)}))Z_1Z_2\right. \\
&\qquad \left. + (\rho_{(p)}^2 - 1)Z_2^2\Big] + 0.5\log\left(4\pi^2(1-\rho^2)\right)\right] \\
=&\frac{-2}{(\rho^2-1)^2}\Big[(\rho_{(p)}^2\rho + \rho - \rho_{(p)} - \rho_{(p)}\rho^2)Z_1^2 + \sqrt{1-\rho_{(p)}^2}(2\rho_{(p)}\rho - \rho^2 - 1)Z_1Z_2 \\
&\qquad + (\rho - \rho_{(p)}^2\rho)Z_2^2\Big] - \frac{\rho}{1-\rho^2}
\end{aligned}$$

Combining this derivative with the objective function (2.31) result in

$$\begin{aligned}
&\frac{\partial}{\partial\rho}B_{OF-\omega}(\boldsymbol{\omega}, \boldsymbol{\omega}^{(p)}) \\
=&K \cdot \sum_{j=1}^{J_k}\sum_{\boldsymbol{Z}\in\Xi} P(\boldsymbol{\xi}|\boldsymbol{Y_{jk}}, \boldsymbol{\omega}^{(p)}, \boldsymbol{\Sigma}^{(p)})\left[\frac{2}{(\rho^2-1)^2}\Big[(\rho_{(p)}^2\rho + \rho - \rho_{(p)} - \rho_{(p)}\rho^2)Z_1^2\right. \\
&\qquad \left. + \sqrt{1-\rho_{(p)}^2}(2\rho_{(p)}\rho - \rho^2 - 1)Z_1Z_2 + (\rho - \rho_{(p)}^2\rho)Z_2^2\Big] + \frac{\rho}{1-\rho^2}\right].
\end{aligned}$$

Since the remaining integral does not depend on the random intercept u, the term $\sum_{k=1}^{K} \sum_{z \in U} P(u|\boldsymbol{Y_k}, \boldsymbol{\omega}^{(p)}, \sigma^{(p)}) = K$ in (2.31) can be introduced as a scalar.

2.4.3 Standard Errors

The standard errors of the estimates are estimated using the bootstrap algorithm. The basic idea behind this algorithm can be described in four steps. First, the observed data are defined as the population. From this population, a new sample is drawn (with replacement) with the same size as the original observed dataset. Second, the same estimation procedure is applied to the new sample and the results are saved. Third, the first two steps are repeated several hundred times, so that several hundred estimates are obtained that all essentially come from different datasets. The last step is to calculate the standard deviation of those estimates, which is the standard error of the estimates. By also calculating the mean of the estimates, confidence intervals can be built as well.

2.4.4 Starting Values

Choosing starting values is a very important part of the estimation of IRT models (and of most models), since poorly chosen starting values can potentially cause the whole estimation to fail or to end up in a local minimum, especially as the model becomes more and more complex. It can also significantly increase the time to converge, which can be especially inconvenient since the EM is known to have problems with slow convergence and to be dependent on starting values.

Speeding up the calculation has been the object of some research over the years in different contexts. Unfortunately, the research on appropriate starting values for an EM algorithm is very thin. This can partly be explained by the variety of different fields in which the EM is used, all of which might need individual solutions. Therefore, there has only been some improvement in specific fields (e.g., in Gaussian mixture modeling: McLachlan (1988), Biernacki, Celeux, and Govaert (2003), Karlis and Xekalaki (2003), and Shireman, Steinley, and Brusco (2015)). However, not much improvement has been achieved for the estimation of latent variable models or IRT models.

Starting Values in Different Software

In current software, the starting values for the latent variable loadings are chosen, for example, as classical test theory estimates (in the software `BILOG` (Du Toit, 2003; M., E., R., & R.D., 2003; Nader, Tran, & Voracek, 2015)), fixed to 1 for all loadings (in the R package `ltm` Rizopoulos (2006)) or as other fixed values depending on the chosen model (e.g., `Mplus`). Another discussed approach is to choose those random values that have the lowest likelihood function value after some initial iterations. Thus, in most cases, the starting values depend on chance, either literally choosing random values or by choosing fixed values and hoping they are close enough even in extreme cases.

Difficulties

Here, some heuristics will be presented for finding appropriate starting values for all parameters. The difficulties δ_i $(i = 1, \dots, I)$ will be chosen as the logit of the empirical probability of solving each item

$$\delta_i = \log\left(\frac{\sum_{j=1}^{N} y_{ij}/N}{1 - \sum_{j=1}^{N} y_{ij}/N}\right),$$

which is the inverse of an IRT model in which only the difficulty is present in the linear predictor (Carlson, 1987).

Loadings

Those loadings $\boldsymbol{\gamma_i}$ are chosen for which the polyserial correlation ρ_{ps} between the dichotomous data and the probability of an IRT model with difficulties (fixed to the starting values) and loadings is the highest

$$\boldsymbol{\gamma_i} = \underset{\hat{\boldsymbol{\gamma}}_{\boldsymbol{i}} \in \mathbb{R}}{\arg\max} \left(\rho_{ps} \left(y_{ijk}, P(Y_{ijk} = 1 | \xi_{jk}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{i}}, \delta_i) \right) \right).$$

Essentially, the model for estimating initial difficulties was simply extended to include the loadings.

To avoid a numerical integration in this step, the latent variables are inserted in the model as fixed parameters (or factor scores). The easiest approach is to calculate the Ordinary Least Square (OLS) estimator for every person, ignoring the multilevel structure:

$$\xi_{jk}^{OLS} = (\boldsymbol{\gamma_i}'\boldsymbol{\gamma_i})^{-1}\boldsymbol{\gamma_i}' y_{ijk}.$$

Since the slope parameters $\boldsymbol{\gamma_i}$ are obviously not available yet, they are substituted with 1. The best results are achieved if the factor scores are built without using items that load (in the model) on more than one latent variable dimension. Finally, the estimated factor scores are scaled, since it is assumed in the model that they follow a standard normal distribution.

More advanced versions could also be chosen here, like Maximum A Posteriori (MAP) estimates. However, since the available model at this point only includes the difficulty and is therefore preliminary in any case, the slight improvement does not justify the computational cost, particularly given that it only serves as an initial value for starting the estimation process of initial values.

Nonlinear Effects

The loadings for the interaction / quadratic terms $\boldsymbol{\Omega_i}$ are chosen in the same way as the loadings of the latent variables $\boldsymbol{\gamma_i}$, but now with the estimated loadings present as:

$$\boldsymbol{\Omega_i} = \underset{\hat{\boldsymbol{\Omega}}_{\boldsymbol{i}} \in \mathbb{R}}{\arg\max} \left(\rho_{ps} \left(y_{ijk}, P(Y_{ijk} = 1 | \boldsymbol{\xi}_{jk}, \boldsymbol{\gamma}_i, \hat{\boldsymbol{\Omega}}_{\boldsymbol{i}}, \delta_i) \right) \right).$$

Naturally, these heuristics work best for many items, particularly when a second dimension of the latent variable is added. Inspired by the use of random values as starting values, these heuristics are improved by initiating them with several random values for the parameters and choosing those for which the observed log-likelihood function of the respective model has the highest value.

One needs to consider that the item parameters should also have limits. If too-high values are chosen, they can lead to items that are solved by everyone or no one, which can potentially cause numerical problems[2]. This needs to be taken into account in the starting values. That is why it is useful to implement ceilings for all parameters that are not exceeded in the estimation, since the results are not interpretable otherwise.

Random Intercept

An initial value for the variance σ^2 of the random intercept in a multilevel model needs to be chosen. There is no existing literature on choosing an appropriate initial value in this context. One heuristic idea that proved to be stable is based on an Analysis of Variance (ANOVA). The variance in the (non)linear predictor depends on the variance that stems from every person

[2] An item that has no variance could lead to values that are calculated to be zero. Since several calculations involve division (e.g., posterior probabilities), this could cause a division by zero.

(the variance of the latent variable ξ_{ij}) and on the variance of the random intercept u_k of every level. In the absence of a multilevel structure, the mean variance of the factor scores (the only source of variance in the (non)linear predictor), which are estimated separately within every cluster, would be the same as the variance of the factor scores estimated using all people. If there is a multilevel structure, the distance between these variances can only stem from the random intercept. Therefore, the initial estimate of the variance σ_u^2 is set to this sum of the differences of the variances. The factor scores are again estimated as before using OLS with the previously estimated parameter values. Since the variances of the latent variables $\boldsymbol{\xi}$ are set to be 1 in every cluster, all variances are accordingly scaled so that the difference properly represents the variance of the random intercept.

Correlation of the Random Variables

In the previously presented estimations of the starting values, the factor scores of the random variables were estimated – for the variance of the random intercept and for the coefficients of the latent variable. The correlation is estimated as the correlation between those most recently estimated factor scores.

Simulations

Naturally, none of these heuristics produce perfect estimates, but they prove to be better than fixing values to 1, for example, as is done in some implementations. In other software several random starting values are drawn and results are calculated for all of them. The results with the best likelihood are then chosen to not end up in a local minimum. Good heuristics may be able to achieve the same result by calculating starting values only once.

All starting values were calculated for clusters of different numbers $N_C = 50, 100, 200$ and sizes $N_S = 50, 100, 150$. Two latent variables were assumed with variances of 1. In the estimation the variances of the latent variables are set to 1 to identify the problem. Ten items were simulated and the true values for the difficulties $\boldsymbol{\delta}$, latent variable coefficients $\boldsymbol{\gamma}$, coefficients of the interactions $\boldsymbol{\Omega}$, the correlation between the latent variables ρ, and the variance of the random intercept σ were set to

$$\begin{aligned}
\boldsymbol{\delta} &= (1, -1.2, -0.2, 0.6, 1.2, -0.6, 0.2, -1, 0, -0.4),\\
\boldsymbol{\gamma} &= \begin{pmatrix} 1 & 0.5 & 0.55 & 1.2 & 0 & 0 & 0 & 0 & 0.45 & 1.1 \\ 0 & 0 & 0 & 0 & 1 & 1.15 & 0.95 & 0.6 & 1.05 & 0.65 \end{pmatrix},\\
\boldsymbol{\Omega}_9 &= \begin{pmatrix} 0 & 0 \\ 0.1 & 0 \end{pmatrix}, \quad \boldsymbol{\Omega}_{10} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},\\
\rho &= 0.3, \quad \text{and}\\
\sigma &= 0.125.
\end{aligned}$$

The biases of the starting values, estimated using the presented methods, are compared to the distance between common fixed values and the true values (‘fix’ in the tables). The results don’t differ between a fixed number of clusters and a fixed size of clusters. Therefore, only the tables for a fixed number of clusters are presented in this section. The results for a fixed size of clusters are given in Appendix B.2 in Tables B.5, B.6, and B.7.

In most software, the difficulties are fixed to 0 in the beginning. The biases in Table 2.1 for the heuristic are very low and in all cases much smaller than the difference between 0 and the true value, except for $\delta_9 = 0$. The variances are very small as well, which indicates that the estimation of the difficulties is stable.

The starting values for latent variable coefficients are often fixed to 1 in software. The biases for $\gamma_{2,1}$, $\gamma_{3,1}$, and $\gamma_{8,2}$ in Table 2.2 are higher compared to the biases of the other loadings. They are also higher than the difference between 1 and the true values, which are not not small with 0.5 and 0.45. The variances for $\gamma_{2,1}$, $\gamma_{3,1}$, and $\gamma_{8,2}$ are high as well. Those item loadings don’t seem to be well approximated by the heuristic.

Table 2.1: RMSE, bias, and variance of the difficulties of simulation of starting values. Data was simulated for model with hierarchical structure ($\sigma^2 = 0.125$) and estimated correlation between the latent variables of $\rho = 0.3$. The number of clusters is fixed to $N_C = 100$. N_S = Number of individuals per cluster. The column 'fix' indicates the difference between the commonly chosen fixed starting value and the true value in the simulation.

N_S	δ_1			δ_2			δ_3			δ_4			δ_5		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.258	-.180	.034	.134	.089	.010	.050	.018	.002	.206	-.143	.022	.318	-.222	.051
100	.253	-.177	.033	.132	.089	.009	.048	.021	.002	.196	-.136	.020	.306	-.215	.047
150	.260	-.183	.034	.133	.091	.009	.043	.018	.002	.202	-.141	.021	.310	-.217	.049
fix		1			-1.2			-.2			.6			1.2	

N_S	δ_6			δ_7			δ_8			δ_9			δ_{10}		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.192	.133	.019	.064	-.035	.003	.149	.100	.012	.041	-.007	.002	.068	.038	.003
100	.195	.136	.020	.058	-.032	.002	.146	.099	.012	.033	-.003	.001	.068	.043	.003
150	.190	.132	.018	.064	-.038	.003	.139	.095	.010	.034	-.008	.001	.065	.041	.003
fix		-.6			.2			-1			0			-.4	

The loadings $\gamma_{9,1}$ and $\gamma_{10,2}$, on the other hand, have a similar true value as $\gamma_{2,1}$, $\gamma_{3,1}$, and $\gamma_{8,2}$, but they are estimated more accurately and with less variance. The biases are much lower than the differences between 1 and the true values.

The biases and variances of the other loadings are in the same order of magnitude. Some biases ($\gamma_{4,1}$, $\gamma_{6,2}$, and $\gamma_{9,2}$) are smaller than or equal to the difference between 1 and the true value, others ($\gamma_{1,1}$, $\gamma_{10,1}$, $\gamma_{5,2}$, and $\gamma_{7,2}$) have a higher bias. Although three of the loadings with higher biases are in items that have true values of 1 and 1.05.

Overall, the results are ambiguous. Adopting the heuristics does not always seem to improve the starting values, but only in a few cases the heuristic produces a bias higher than the difference between 1 and the true values.

The interaction effects are usually set to 0 in the beginning. The heuristic performs much better for both, the small true value of 0.1 and for the higher value of 1, than the fixed value. the variances are also small and indicate a

Table 2.2: RMSE, bias, and variance of loadings of the latent variables of simulation of starting values. Data was simulated for model with hierarchical structure ($\sigma^2 = 0.125$) and estimated correlation between the latent variables of $\rho = 0.3$. The number of clusters is fixed to $N_C = 100$. N_S = Number of individuals per cluster. The column 'fix' indicates the difference between the commonly chosen fixed starting value and the true value in the simulation.

N_S	$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.283	.199	.040	.835	.590	.349	.868	.613	.377	.140	.096	.010
100	.278	.196	.039	.838	.593	.352	.864	.611	.374	.134	.093	.009
150	.281	.198	.040	.837	.592	.350	.868	.614	.377	.133	.093	.009
fix		0			.5			.45			-.2	

N_S	$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.214	.145	.025	.391	-.273	.078	.297	.210	.044	.226	.158	.026
100	.204	.142	.022	.391	-.275	.077	.298	.210	.045	.222	.156	.025
150	.201	.140	.021	.400	-.282	.081	.300	.212	.045	.222	.156	.025
fix		.55			-.1			0			-.15	

N_S	$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.498	.351	.125	.682	.482	.233	.108	.060	.008	.099	-.056	.007
100	.496	.350	.123	.687	.485	.236	.098	.061	.006	.102	-.065	.006
150	.494	.349	.123	.686	.485	.236	.090	.058	.005	.095	-.063	.005
fix		.05			.4			-.05			.35	

Table 2.3: RMSE, bias, and variance of the interaction coefficients of simulation of starting values. Data was simulated for model with hierarchical structure ($\sigma^2 = 0.125$) and estimated correlation between the latent variables of $\rho = 0.3$. The number of clusters is fixed to $N_C = 100$. N_S = Number of individuals per cluster. The column 'fix' indicates the difference between the commonly chosen fixed starting value and the true value in the simulation.

N_S	$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$			ρ			σ^2		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.153	.017	.023	.477	-.327	.121	.201	-.142	.020	.043	-.027	.001
100	.105	.002	.011	.468	-.325	.114	.202	-.143	.021	.031	-.018	.001
150	.078	.001	.006	.474	-.331	.115	.199	-.141	.020	.025	-.012	.000
fix		.1			1			.3			.125	

stable estimation.

Unsurprisingly, the heuristics for the correlation and the variance of the random intercept are better than setting them to 0. Especially the variance σ^2 is estimated accurately in this simulation and with a very low variance.

It is neither surprising nor worrying that an increasing number of clusters or cluster sizes does not improve the calculated starting values. It is of more importance to consistently choose good starting values rather than have consistent starting values.

2.4.5 Convergence

In Section 2.3.4 it was established that the MINoLEM estimation process is a twofold iteration. On the one hand, the classical E-step and M-step are iterated, and on the other hand, the estimation of the item parameters and of the variance of the random intercept are iterated within the M-step. To establish a convergence criterion one needs to observe the following: the objective function changes within each of these iterations. This is due to the posterior probabilities, which are built with estimates from the respective previous iteration. As a result, not only can the distinct objective functions for the item parameters (2.19) and the variance of the random intercept (2.20) not be compared, neither also not the objective functions for either the item parameters or the variance of the random intercept between two M-steps. The value of the objective function may even decrease between two iterations. Therefore, the observed data log-likelihood

$$\begin{aligned}&L_{\text{obs}}(\boldsymbol{Y}|\boldsymbol{\omega}^{(p)},\sigma^{(p)})\\=&\sum_{k=1}^{K}\log\left(\sum_{u\in U}\prod_{j=1}^{J_k}\sum_{\boldsymbol{\xi}\in\Xi}L(\boldsymbol{Y_{jk}}|\boldsymbol{\xi},u,\boldsymbol{\omega}^{(p)},\sigma^{(p)})P(\boldsymbol{\xi}|\boldsymbol{\Sigma_{\xi}})P(u|\sigma^{(p)})\right)\end{aligned}$$

needs to be calculated in each M-step to measure the rate of convergence.

The stop criterion is then given by the distance between the last two evaluations of L_{obs} after the estimation of σ^2 divided by the distance between the last evaluation of the observed log-likelihood and the evaluation of L_{obs} after the estimation of the starting values, to be less dependent on the scale of the likelihood:

$$\frac{L_{\text{obs}}(\boldsymbol{Y}|\boldsymbol{\omega}^{(p+1)}, \sigma^{(p+1)}) - L_{\text{obs}}(\boldsymbol{Y}|\boldsymbol{\omega}^{(p)}, \sigma^{(p)})}{L_{\text{obs}}(\boldsymbol{Y}|\boldsymbol{\omega}^{(p+1)}, \sigma^{(p+1)}) - L_{\text{obs}}(\boldsymbol{Y}|\boldsymbol{\omega}^{(0)}, \sigma^{(0)})} < \epsilon.$$

In contrast to the objective functions of the EM, the observed log-likelihood is monotonous, as shown in Section 1.6, so that every improvement of the parameter estimates is mirrored by an increase in the observed log-likelihood function.

2.4.6 Numerical Accuracy

The more people there are in a cluster, the more probability terms need to be multiplied to form the posterior probabilities of the random intercept. However, a computer cannot handle arbitrarily small values and R has a specific limit for the smallest representable number, which depends on the machine R is installed on. This may lead to products of probabilities being rounded to 0 and therefore potentially to infinite or NaN values if there is a division by 0. To avoid this problem, the accuracy of R needs to be increased in these cases, which can be achieved using the R package `Rmpfr` by Maechler (2020).

The accuracy of the estimation is also influenced by the chosen optimization method in the M-step. A common approach for nonlinear problems are quasi-Newton methods. Here, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm was applied, which is implemented in the `optim` function in R.

Chapter 3

Data Simulations

In this chapter, the proposed MINoLEM estimation is tested in several simulation studies and compared to other implementations. As a part of this, a selection of software solutions for IRT models is presented and their differences are briefly discussed. Subsequently, data without hierarchical structure are simulated to investigate the performance of MINoLEM for different model formulations. Finally, data for the full model, as presented in Equation (2.1), are simulated, and the results are discussed.

All deductions in Chapter 2 are valid for quadratic and interaction effects of the latent variables. In the implementation, however, interaction effects were prioritized, since interaction effects are considered more complex than quadratic effects.

3.1 Different Implementations for IRT Models

Various latent variable frameworks were introduced in Chapter 1. The estimation of IRT models is, to a certain extent, possible in all of these frameworks. Therefore, different implementations exist that are able to estimate IRT models.

Table 3.1: Common IRT models that can be estimated by each program. 'Multivariate' indicates that the software is able to estimate latent variables with more than one dimension. 'Multilevel' indicates that the software is able to estimate models which include a hierarchical structure of the observed data. 'Interaction' indicates that the software is able to estimate models that incorporate nonlinear latent variable effects as defined in Section 1.2.3. ‡ only sirt.

Model	mirt	lavaan	TAM	Mplus	GLLAMM	flexmirt	ltm	MINoLEM
1PL/Rasch	✓	✓	✓	✓	✓	✓	✓	✓
2PL	✓	✓	✓	✓	✓	✓	✓	✓
3PL	✓		✓			✓	✓	
4PL	✓							
Part. Credit	✓	✓	✓	✓	✓	✓	✓	
Gr. Response	✓	✓		✓	✓	✓	✓	
Multivariate	✓	✓	✓	✓	✓	✓	✓	✓
Multilevel	✓		✓[‡]	✓	✓	✓		✓
Interaction							✓	✓

In this section, three different sources are considered

- R packages (mirt (Chalmers, 2012), lavaan (Rosseel, 2012), TAM (Robitzsch, Kiefer, & Wu, 2020), and ltm (Rizopoulos, 2006)),
- GLLAMM implemented in Stata (StataCorp., 2019), and
- paid programs (Mplus (L. Muthén & Muthén, 1998-2017) and flexmirt (Cai, 2017)).

The implementations of mirt, TAM, and flexmirt, are based on the classical IRT framework, while lavaan and Mplus estimate IRT from the point of view of SEM. GLLAMM and ltm are based on generalized linear models (see Section 1.4 and Section 2.3.4). A wider range of R packages are inspected in Choi and Asilkalkan (2019), and a comparison of commercial software can be found in Han and Paek (2014).

Table 3.2: Implemented estimation methods for dichotomous dependent variables in each program. The estimation procedure available for the most complex models possible in each respective software (thus including multilevel or nonlinear latent effects) is circled.

Estimator	`mirt`	`lavaan`	`sirt` / `TAM`	`Mplus`	`GLLAMM`	`flexmirt`	`ltm`
MLE	✓		✓	✓	✓	✓	✓
EM	✓		✓	✓		✓	(✓)
JML			✓	✓			
WLSMV		(✓)		(✓)			
MCMC			(✓)	✓			
MHRM	(✓)					(✓)	

3.1.1 Range of Implementations

Table 3.1 provides an overview of the different IRT models, which the aforementioned programs are able to estimate. All considered implementations cover the 1PL, 2PL, and PCM, and all can handle multivariate latent variables. With the exception of `ltm` and `lavaan`[1], all packages can include a multilevel structure, but `ltm` is the only program that can handle a nonlinear influence of the latent variables. Overall, the `mirt` package is the most capable IRT software in this regard and is also one of the most widely used implementations. However, there is no software that can simultaneously incorporate nonlinear latent variable effects and a multilevel data structure. The MINoLEM introduced in this thesis cannot yet be applied to all IRT models, but it expands the range of applicable models.

3.1.2 Different Estimators

As described in Section 1.5, not only different frameworks but also different estimation procedures can be applied. Table 3.2 shows some of the estimation procedures, which the different software can apply, if dichotomous observed data are present.

[1] Generally `lavaan` can handle hierarchical data, but not for categorical dependent variables.

An MLE, which is in most cases implemented using the EM algorithm, is applied by all software except lavaan. JML is implemented in the TAM and Mplus software, because the number of items in large scale assessments is usually very high and its lack of consistency might not be a problem in these cases (see e.g., Haberman (1977) for Rasch-models).

The WLSMV estimator is implemented in lavaan and Mplus. It is worth noting that both implementations use a probit link and that the parameters are given in SEM notation. To get IRT notation with a logit link, a transformation is necessary (see Section 1.4.3).

The MHRM algorithm is implemented in the software flexmirt (which is co-created by the author of MHRM) and in mirt, while MCMC methods are possible in TAM and Mplus.

The estimation procedure available for the most complex model in each respective software (thus including multilevel or nonlinear latent effects) is circled in the table.

3.2 Simulation Study – Single-Level

In the first simulation study, the estimation procedure is analyzed for a model without hierarchical structure (without random intercept). The results will be compared to the implementation by Rizopoulos and Moustaki (2008) in the R package ltm (Rizopoulos, 2006), which is able to estimate nonlinear latent variable effects in a single-level context. The model to simulate the data is given by

$$P(Y_{ij} = 1|\boldsymbol{\xi}_j, \boldsymbol{\gamma_i}, \boldsymbol{\Omega_i}, \delta_i) = \frac{1}{1 + \exp(-(\boldsymbol{\xi}_j'\boldsymbol{\gamma}_i + \boldsymbol{\xi}_j'\boldsymbol{\Omega}_i\boldsymbol{\xi}_j - \delta_i))}.$$

The difficulties and the latent variable coefficients are set to

$$\boldsymbol{\delta} = (1, -1.2, -0.2, 0.6, 1.2, -0.6, 0.2, -1, 0, -0.4) \quad \text{and} \tag{3.1}$$

$$\boldsymbol{\gamma} = \begin{pmatrix} 1 & 0.5 & 0.55 & 1.2 & 0 & 0 & 0 & 0 & 0.45 & 1.1 \\ 0 & 0 & 0 & 0 & 1 & 1.15 & 0.95 & 0.6 & 1.05 & 0.65 \end{pmatrix}^t , \tag{3.2}$$

respectively. The latent variable $\boldsymbol{\xi}$ is assumed to be two-dimensional and follow the multivariate standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I_2})$. Its variances are set to 1 to identify the model. The last two items are influenced by the interaction of the latent variables. An interaction however is most meaningful if the item is also influenced by the main effects of the corresponding latent variables. Therefore, both items are set to have cross-loadings. Furthermore, defining one of those items to be dependent on only one main effect, could cause problems in the estimation of both, the interaction and the main effects.

3.2.1 Uncorrelated Latent Variables

Unfortunately, the correlation cannot be estimated in the `ltm` package. Since the correlation is considered an important aspect in this thesis, the simulation is separated into two parts – first without and then with correlation. Furthermore, two different sets of coefficients for the latent interaction – moderate and high – are examined. The number of observations was varied between $N = 500, 1000, 2000, 5000$ and 200 datasets were simulated for each condition.

Moderate Interaction Effects

First, moderate interactions are investigated for the last two items with

$$\boldsymbol{\Omega}_9 = \begin{pmatrix} 0 & 0 \\ 0.1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega}_{10} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} . \tag{3.3}$$

Table 3.3: RMSE, bias, and variance of the difficulties of the last five items for simulation of single-level model and fixed correlation between the latent variables to $\rho = 0$. The interaction terms have medium values. E_M = Estimation with MINoLEM. E_L = Estimation with `ltm`.

N		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.145	.026	.020	.123	-.019	.015	.099	-.001	.010	.125	-.003	.016	.157	.021	.024
	E_L	.147	.028	.021	.124	-.019	.015	.099	.001	.010	.133	.001	.018	.159	.025	.025
1000	E_M	.099	.006	.010	.086	-.011	.007	.066	-.002	.004	.089	.015	.008	.099	.008	.010
	E_L	.100	.004	.010	.085	-.010	.007	.066	-.002	.004	.092	.017	.008	.101	.009	.010
2000	E_M	.073	.014	.005	.061	.001	.004	.050	-.002	.002	.067	.008	.004	.075	.007	.006
	E_L	.074	.012	.005	.061	.002	.004	.050	-.001	.002	.069	.009	.005	.075	.006	.006
5000	E_M	.043	.003	.002	.035	-.001	.001	.031	-.004	.001	.042	-.002	.002	.044	-.000	.002
	E_L	.042	.000	.002	.035	-.000	.001	.031	-.004	.001	.042	-.003	.002	.045	-.001	.002

N_C		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.131	.008	.017	.112	.020	.012	.112	-.004	.013	.164	.018	.027	.162	.013	.026
	E_L	.140	.001	.020	.109	.020	.012	.112	-.008	.013	.140	.020	.019	.196	-.016	.038
1000	E_M	.083	.004	.007	.074	.005	.005	.076	-.003	.006	.091	.001	.008	.115	.014	.013
	E_L	.084	.003	.007	.073	.004	.005	.075	-.002	.006	.091	.002	.008	.130	-.007	.017
2000	E_M	.058	-.004	.003	.055	-.001	.003	.060	-.001	.004	.059	.002	.004	.070	.020	.005
	E_L	.058	-.004	.003	.055	-.002	.003	.060	-.001	.004	.060	.004	.004	.075	.001	.006
5000	E_M	.043	.002	.002	.033	-.003	.001	.034	.000	.001	.039	-.005	.002	.043	.010	.002
	E_L	.043	.002	.002	.033	-.003	.001	.034	.000	.001	.039	-.003	.002	.050	-.010	.002

The biases of the difficulties in Table 3.3 are low and almost equal for `ltm` (E_L) and MINoLEM (E_M). They tend to decline with a rising number of individuals, which indicates consistent estimation of the difficulties. The variances decrease in all cases, as does the RMSEs. For all sizes of the dataset, the estimation seems to be very stable, and with a small dataset of $N = 500$, the estimation with `ltm` or MINoLEM produces only a small bias. With $N = 5000$, the biases and variances are very low, and the estimation is very accurate and efficient.

The biases of the coefficients of the latent variables in Table 3.4 are in the same order of magnitude as the biases of the difficulties. The estimates of the cross-loadings do not have an increased bias, but $\gamma_{10,1}$ has a higher variance in `ltm` and MINoLEM for $N = 500$, which indicates that some datasets are

Table 3.4: RMSE, bias, and variance of the coefficients of the first latent dimension for simulation of single-level model and fixed correlation between the latent variables to $\rho = 0$. The interaction terms have medium values. E_M = Estimation with MINoLEM. E_L = Estimation with `ltm`.

N		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.267	.069	.066	.186	.004	.035	.180	.012	.032	.311	.013	.097
	E_L	.267	.065	.067	.184	.001	.034	.178	.005	.032	.352	.039	.123
1000	E_M	.176	.024	.030	.137	.008	.019	.115	.003	.013	.200	.023	.039
	E_L	.183	.019	.033	.137	.003	.019	.118	-.002	.014	.233	.034	.053
2000	E_M	.140	.032	.018	.089	.006	.008	.086	.009	.007	.149	.022	.022
	E_L	.145	.027	.020	.088	.002	.008	.084	.003	.007	.166	.025	.027
5000	E_M	.074	.007	.005	.058	.007	.003	.058	.014	.003	.089	.009	.008
	E_L	.074	-.000	.006	.057	.004	.003	.056	.010	.003	.095	.007	.009

N		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.247	.049	.059	.451	.013	.203	.256	.016	.065	.287	.037	.010
	E_L	.212	.034	.044	.503	.088	.246	.265	.025	.069	.323	.058	.010
1000	E_M	.147	.016	.021	.257	-.009	.066	.169	.007	.029	.167	.006	.010
	E_L	.147	.016	.021	.312	.053	.095	.177	.011	.031	.172	.009	.010
2000	E_M	.087	.002	.008	.179	-.065	.028	.116	.006	.013	.119	.014	.010
	E_L	.088	.003	.008	.191	-.007	.036	.117	.006	.014	.121	.014	.010
5000	E_M	.058	-.002	.003	.118	-.041	.012	.078	-.002	.006	.078	.006	.010
	E_L	.058	-.000	.003	.134	.024	.017	.079	-.003	.006	.080	.004	.010

N		$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.233	.002	.054	.163	-.001	.027	.456	.147	.186	.255	.006	.065
	E_L	.239	.003	.057	.163	-.007	.026	.292	.099	.076	.317	.058	.097
1000	E_M	.158	-.001	.025	.124	.002	.015	.186	.051	.032	.200	.004	.040
	E_L	.159	-.005	.025	.124	.000	.015	.184	.048	.032	.231	.040	.052
2000	E_M	.117	.016	.013	.088	.002	.008	.120	.021	.014	.116	-.029	.013
	E_L	.116	.012	.013	.088	.001	.008	.121	.020	.014	.122	.001	.015
5000	E_M	.068	.004	.005	.054	.001	.003	.066	.006	.004	.076	-.019	.005
	E_L	.067	.001	.005	.054	.000	.003	.067	.005	.004	.085	.013	.007

not as well estimated as most.

Especially for smaller datasets, the estimation can converge to a local minimum, which is far away from the true values. The MINoLEM approach seems to have fewer of those cases, and this problem occurs less likely with an increasing size of the datasets. To counter this problem in a reasonable way, a dataset was not taken into account (also in the following simulations), if at least one parameter estimate was estimated above 5 in absolute values. For those, it was assumed that the estimation diverged and that it would also be apparent in an applied case. Nevertheless, the higher variance can be explained by remaining poorly estimated datasets.

Overall, the variances are small and decline with increasing number of individuals. The RMSEs decline as well, and are mostly in the same order of magnitude as for the difficulties. The interaction effects in Table 3.5 show slightly higher biases than the difficulties and coefficients for both methods. The biases of $\Omega_{10}^{(2,1)}$ do not seem to decline with more individuals in MINoLEM. The variances decrease, and while the first interaction effect $\Omega_{9}^{(2,1)}$ is estimated as well as most factor loadings concerning the RMSEs, the RMSEs of the second interaction effect $\Omega_{10}^{(2,1)}$ and of the factor loading $\gamma_{10,1}$ are higher. This can be explained by the relatively high true interaction coefficient, as detailed in the following paragraph.

Table 3.5: RMSE, bias, and variance of the interaction coefficients for simulation of single-level model and fixed correlation between the latent variables to $\rho = 0$. The interaction terms have medium values. E_M = Estimation with MINoLEM. E_L = Estimation with `ltm`.

N		$\Omega_{9}^{(2,1)}$			$\Omega_{10}^{(2,1)}$		
		rmse	bias	var	rmse	bias	var
500	E_M	.521	.058	.268	.537	-.015	.288
	E_L	.428	.234	.128	.712	.127	.491
1000	E_M	.233	.027	.054	.416	-.025	.172
	E_L	.214	.104	.035	.479	.077	.223
2000	E_M	.169	.014	.028	.239	-.072	.052
	E_L	.141	.052	.017	.275	.024	.075
5000	E_M	.099	.020	.009	.168	-.063	.024
	E_L	.080	.013	.006	.195	.042	.036

The influence of the interaction effect on the Item Response Surface (IRS) can be seen in Figures 3.1 to 3.4. Each plot shows the probability of answering an item correctly, depending on latent variables η_1 and η_2. The

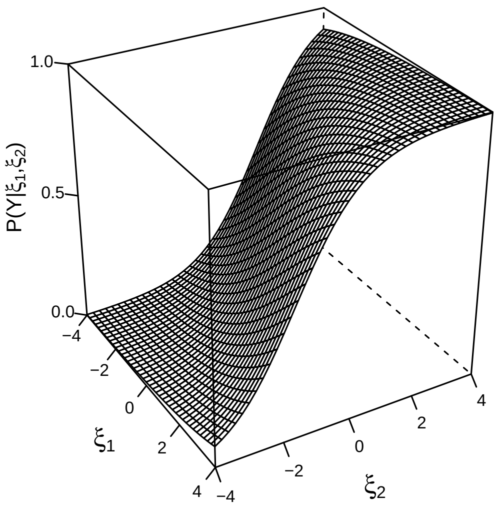

Figure 3.1: Item response surface of 2-dimensional IRT model item without interaction effect.

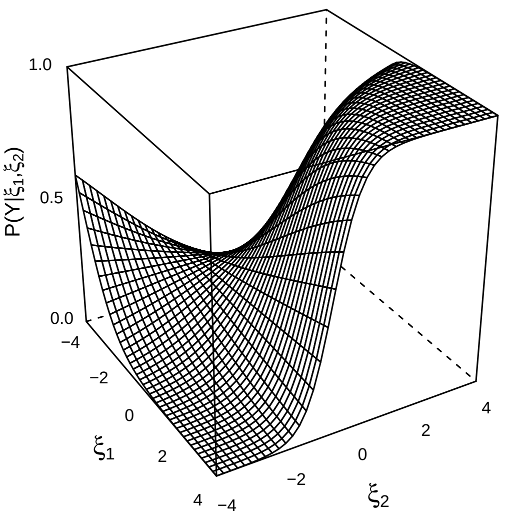

Figure 3.2: Item response surface of 2-dimensional IRT model item with low interaction effect of $\Omega = 0.4$.

interaction coefficient is varied while the other item parameters are fixed to the values of item nine of the previous simulation.

The IRS of an item without an interaction effect, as in Figure 3.1, is a bent surface, where the degree of the bend depends solely on the coefficients of the latent variables. Adding an interaction coefficient of $\Omega = 0.4$ introduces a twist in the surface and hints at a saddle (see Figure 3.2). Figure 3.3 with $\Omega = 1.6$ shows that further increase of the interaction coefficient creates a saddle, which gets narrower with higher interaction coefficients.

With rising difficulty of the item, the saddle becomes deeper. The coefficients of the individual latent variables determine where the saddle lies in the ξ_1-ξ_2-plane. Introducing a negative interaction coefficient instead, turns the orientation of the saddle by 90°.

The differences between the Figures 3.1 and 3.2 and the Figures 3.3 and 3.4 each lies in an increase of the interaction coefficient by 0.4. However, the

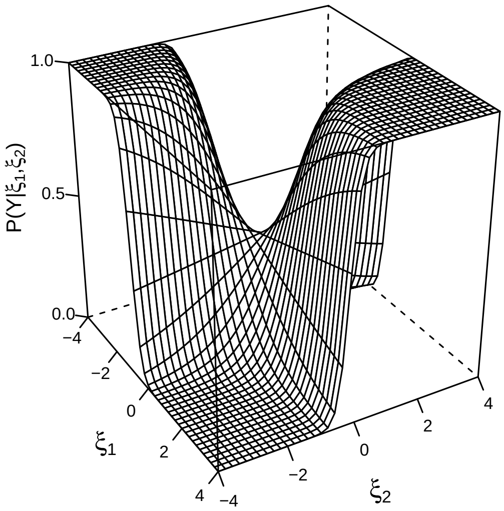

Figure 3.3: Item response surface of 2-dimensional IRT model item with moderate interaction effect of $\Omega = 1.6$.

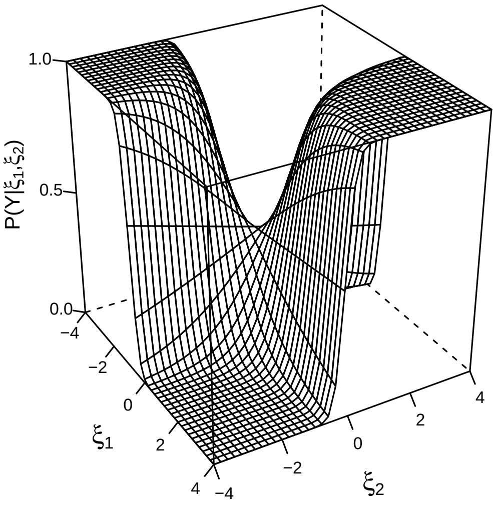

Figure 3.4: Item response surface of 2-dimensional IRT model item with high interaction effect of $\Omega = 2$.

distinction is much more apparent between Figures 3.1 and 3.2, while Figures 3.3 and 3.4 only show a slight difference, if looked closely. This explains why especially the efficient estimation (with low standard error) of high interaction coefficients is more difficult than for low interaction coefficients.

This can also be seen in Table 3.5, where the first interaction coefficient is set to $\Omega_9^{(2,1)} = 0.1$, while the second is set to $\Omega_{10}^{(2,1)} = 1$. The RMSE is greater for $\Omega_{10}^{(2,1)}$, which is coherent with the previous explanation, since distinctions get more difficult between different coefficients of latent interactions, when they increase in size.

High Interaction Effects

Setting both interaction effects to high values

$$\mathbf{\Omega}_9 = \begin{pmatrix} 0 & 0 \\ -1.9 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{\Omega}_{10} = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}, \tag{3.4}$$

while keeping all other parameters, demonstrates even better that the estimation of high interaction effects is difficult. The number of observations was again varied between N= $500, 1000, 2000, 5000$ and 200 datasets were simulated for each condition.

Table 3.6: RMSE, bias, and variance of the difficulties of the last five items for simulation of single-level model and fixed correlation between the latent variables to $\rho = 0$. The interaction terms have high values. E_M = Estimation with MINoLEM. E_L = Estimation with `ltm`.

N		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.139	.024	.019	.124	-.024	.015	.097	.008	.009	.130	.008	.017	.142	.016	.020
	E_L	.141	.025	.019	.124	-.019	.015	.098	.002	.010	.127	-.002	.016	.152	.020	.023
1000	E_M	.096	.002	.009	.089	-.014	.008	.067	-.004	.004	.091	.017	.008	.099	.004	.010
	E_L	.097	.004	.009	.086	-.012	.007	.066	-.002	.004	.089	.013	.008	.101	.010	.010
2000	E_M	.071	.012	.005	.061	.002	.004	.049	-.001	.002	.065	.006	.004	.076	.011	.006
	E_L	.069	.009	.005	.060	.003	.004	.050	-.001	.002	.067	.006	.004	.076	.008	.006
5000	E_M	.040	.004	.002	.036	-.002	.001	.031	-.004	.001	.041	.001	.002	.045	.004	.002
	E_L	.040	.000	.002	.036	-.000	.001	.031	-.003	.001	.042	-.002	.002	.044	.001	.002
N_C		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.130	.002	.017	.113	.026	.012	.109	-.008	.012	.186	.025	.034	.186	.003	.035
	E_L	.133	.006	.018	.109	.020	.011	.114	-.009	.013	.191	.007	.037	.978	-.146	.935
1000	E_M	.083	.006	.007	.075	.010	.006	.072	-.001	.005	.106	.009	.011	.118	.021	.014
	E_L	.084	.004	.007	.074	.005	.005	.073	-.000	.005	.112	.005	.012	.174	-.014	.030
2000	E_M	.060	.000	.004	.055	.000	.003	.060	.002	.004	.088	.011	.008	.106	.017	.011
	E_L	.057	-.003	.003	.055	-.002	.003	.060	.000	.004	.076	-.005	.006	.106	-.007	.011
5000	E_M	.043	.003	.002	.034	-.001	.001	.034	.001	.001	.058	-.003	.003	.076	.026	.005
	E_L	.042	.002	.002	.033	-.004	.001	.034	.000	.001	.050	-.010	.002	.059	-.006	.003

First, it can be said that the estimation of the difficulties in Table 3.6 is not affected by the higher interaction values. The bias is as low as for medium interaction values and declines as well. The RMSEs and variances are also in the same order of magnitude.

Table 3.7: RMSE, bias, and variance of the coefficients of the first latent dimension for simulation of single-level model and fixed correlation between the latent variables to $\rho = 0$. The interaction terms have high values. E_M = Estimation with MINoLEM. E_L = Estimation with `ltm`.

N		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.254	.060	.061	.172	.005	.029	.173	.024	.029	.312	.039	.096
	E_L	.246	.058	.057	.178	.003	.032	.180	.001	.032	.346	.027	.119
1000	E_M	.172	.021	.029	.135	.012	.018	.114	.001	.013	.180	.017	.032
	E_L	.170	.019	.028	.133	.011	.018	.114	-.003	.013	.196	.012	.038
2000	E_M	.133	.029	.017	.086	.006	.007	.084	.008	.007	.152	.013	.023
	E_L	.129	.018	.016	.084	.002	.007	.082	.004	.007	.155	.014	.024
5000	E_M	.074	.009	.005	.061	.008	.004	.055	.011	.003	.096	.022	.009
	E_L	.067	.001	.004	.058	.004	.003	.052	.006	.003	.090	.009	.008

N		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.432	.050	.184	.483	-.037	.232	.241	.012	.058	.295	.060	.010
	E_L	.963	.163	.901	3.459	.920	11.118	.250	.013	.062	.291	.036	.010
1000	E_M	.309	.016	.095	.405	-.057	.161	.164	.008	.027	.182	.017	.010
	E_L	.295	.031	.086	1.447	.276	2.019	.168	.016	.028	.180	.009	.010
2000	E_M	.289	.042	.082	.327	-.099	.097	.122	.010	.015	.128	.013	.010
	E_L	.180	.010	.032	.314	.035	.097	.123	.010	.015	.119	.007	.010
5000	E_M	.220	-.003	.049	.274	-.112	.063	.081	.004	.007	.084	.012	.010
	E_L	.115	-.015	.013	.186	.035	.033	.080	.003	.006	.076	.004	.010

N		$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.221	.009	.049	.162	-.001	.026	.504	.051	.251	.345	.023	.118
	E_L	.223	.011	.050	.165	-.003	.027	4.650	1.229	2.116	1.847	.379	3.268
1000	E_M	.158	.013	.025	.114	-.003	.013	.482	.061	.228	.310	-.019	.096
	E_L	.154	-.000	.024	.115	-.006	.013	1.059	.224	1.072	.384	.057	.144
2000	E_M	.115	.019	.013	.087	-.002	.008	.415	.066	.168	.260	-.049	.065
	E_L	.113	.014	.012	.087	-.003	.008	.737	.123	.529	.213	-.004	.046
5000	E_M	.070	.006	.005	.054	.002	.003	.310	-.013	.096	.185	-.063	.030
	E_L	.068	-.001	.005	.053	-.001	.003	.170	.022	.028	.125	.003	.016

The coefficients of the latent variables of the items, that are only affected by one main effect ($\gamma_{1,1}, \gamma_{2,1}, \gamma_{3,1}, \gamma_{4,1}, \gamma_{5,2}, \gamma_{6,2}, \gamma_{7,2}$, and $\gamma_{8,2}$) in Table 3.7, are also not affected in both methods as well. RMSEs, biases, and variances are as low as before, and the biases are mostly decreasing and indicating consistency.

However, the cross-loadings in `ltm` are not being estimated adequately in some cases. There seem to be more datasets for which `ltm` ‘diverges’, although datasets with any estimated values of 5 and above are not considered again. MINoLEM does not show this behavior and estimates the cross-loadings stable for all sizes of the dataset.

Nevertheless, the values do not decrease as much as for the medium interaction effects in the previous section. The explanation is given above as well by the increased complexity of estimating interaction effects and the respective loadings if they values get higher.

Table 3.8: RMSE, bias, and variance of the interaction coefficients for simulation of single-level model and fixed correlation between the latent variables to $\rho = 0$. The interaction terms have high values. E_M = Estimation with MINoLEM. E_L = Estimation with `ltm`.

N		$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$		
		rmse	bias	var	rmse	bias	var
500	E_M	.823	-.065	.674	.757	-.052	.571
	E_L	6.704	-1.931	41.220	5.629	1.652	28.960
1000	E_M	.794	-.049	.627	.719	-.062	.513
	E_L	1.506	-.323	2.164	2.051	.463	3.991
2000	E_M	.754	-.094	.560	.656	-.107	.419
	E_L	1.104	-.193	1.181	.588	.119	.332
5000	E_M	.631	.011	.398	.494	-.215	.198
	E_L	.291	-.036	.084	.327	.054	.104

The results in Table 3.8 for the interaction effects confirm this observation. Again, `ltm` has various issues in the estimation, especially for smaller datasets, while MINoLEM is stable in all conditions. For $N = 5000$, `ltm` does not show these estimation issues anymore, but the values stay slightly higher than before, when smaller interaction effects were estimated.

Nonetheless, the simulations show that the estimation of interaction effects, difficulties, and coefficients is handled well by `ltm` and slightly more stable by MINoLEM. Low biases and variances could be observed in all conditions.

3.2.2 Correlated Latent Variables

In this simulation, the correlation between the latent variables, set to $\rho = 0.3$, is estimated as well. The `ltm` package cannot estimate correlations and fixes it to $\rho = 0$. The misspecification results in high RMSE values. Therefore, only the results of the MINoLEM approach are presented.

Since the focus lies on the interaction and the correlation between the latent variables, only those results are given here. The results of the difficulties and the coefficients of the latent variables are not affected by the additional estimation of the correlation and are given in Appendix B.1 in Tables B.1 to B.4.

Table 3.9: RMSE, bias, and variance of the interaction coefficients for simulation of single-level model and estimated correlation between the latent variables of $\rho = 0.3$. The interaction terms have medium values. E_M = Estimation with MINoLEM.

N		$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$			ρ		
		rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.429	-.015	.184	.570	.095	.316	.123	-.014	.015
1000	E_M	.280	.018	.078	.360	-.068	.125	.090	-.028	.007
2000	E_M	.204	-.003	.042	.218	-.039	.046	.054	-.014	.003
5000	E_M	.124	.019	.015	.175	-.058	.027	.048	-.023	.002

Tables 3.9 and 3.10 show that the additional estimation of the correlation between the latent variables only marginally influences the estimation of the interaction terms for medium and high values, as defined in Equations (3.3) and (3.4). In comparison to the simulation, in which the correlation was fixed to $\rho = 0$, the biases are slightly higher, while the variances are lower. At first sight, estimation of the correlation appears to be more efficient, but it also slightly increases the bias.

Table 3.10: RMSE, bias, and variance of the interaction coefficients for simulation of single-level model and estimated correlation between the latent variables of $\rho = 0.3$. The interaction terms have high values. E_M = Estimation with MINoLEM.

N		$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$			ρ		
		rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.636	.186	.370	.657	-.261	.364	.112	-.028	.012
1000	E_M	.465	.185	.182	.676	-.374	.317	.085	-.035	.006
2000	E_M	.421	.207	.134	.519	-.301	.179	.057	-.023	.003
5000	E_M	.355	.196	.088	.506	-.320	.153	.053	-.029	.002

The correlation is estimated with a low bias, which is not influenced by increasing the value of the interaction effects. The bias does not decrease with an increasing number of individuals and is negative in all cases, which indicates a slight underestimation. The variances and therefore the RMSEs decrease with more individuals in the datasets and have values as low as those of the difficulties in the previous simulation.

3.2.3 Estimation of Zero Loadings

In this simulation, the difficulties and loadings are, as before, defined in Equations (3.1) and (3.2). As in Equation (3.3), the interaction effects take moderate values, and the latent variables are assumed to be uncorrelated, to allow the comparison with `ltm`.

Contrary to the estimation in Section 3.2.1, the coefficient $\gamma_{8,1}$ of the reading ability and the loading $\Omega_8^{(2,1)}$ of an interaction effect for item eight are freely estimated, instead of being fixed to their true value 0. This tests the handling of the wrong assumption that item eight does not only depend on the second latent variable dimension but also on the first dimension, as well as on the interaction of both latent variables.

The results show that MINoLEM and `ltm` have no problem detecting the true relationships. The biases, variances, and RMSEs of the difficulty of item eight in Table 3.11 are basically unchanged in comparison to Section 3.2.1 with slightly higher values. The results of the other item difficulties are not affected.

Table 3.11: RMSE, bias, and variance of difficulties of simulation for single-level model and correlation between the latent variables of 0. The loading $\gamma_{8,1}$ of the reading ability and the loading $\Omega_8^{(2,1)}$ of an interaction effect for item 8 are simulated to be 0 but estimated nevertheless.

N_C		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.144	.028	.020	.123	-.020	.015	.098	-.000	.010	.130	.001	.017	.158	.023	.024
	E_L	.183	.033	.032	.123	-.019	.015	.098	.001	.010	.137	.003	.019	.162	.027	.025
1000	E_M	.101	.007	.010	.086	-.011	.007	.067	-.001	.004	.088	.015	.008	.100	.009	.010
	E_L	.101	.005	.010	.085	-.010	.007	.066	-.002	.004	.092	.017	.008	.101	.009	.010
2000	E_M	.073	.014	.005	.061	.001	.004	.050	-.002	.002	.067	.008	.004	.075	.007	.006
	E_L	.074	.012	.005	.061	.003	.004	.050	-.001	.002	.069	.008	.005	.075	.006	.006
5000	E_M	.043	.003	.002	.035	-.001	.001	.031	-.004	.001	.042	-.002	.002	.044	-.000	.002
	E_L	.043	.000	.002	.035	-.000	.001	.031	-.004	.001	.043	-.002	.002	.044	-.001	.002

N_C		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.136	.003	.018	.112	.021	.012	.131	-.026	.016	.137	.015	.018	.160	.018	.025
	E_L	.141	.001	.020	.110	.021	.012	.135	-.031	.017	.139	.020	.019	.201	-.018	.040
1000	E_M	.084	.003	.007	.074	.005	.005	.079	-.012	.006	.092	-.001	.008	.112	.020	.012
	E_L	.084	.003	.007	.074	.004	.005	.080	-.012	.006	.091	.002	.008	.132	-.008	.017
2000	E_M	.058	-.004	.003	.055	-.001	.003	.061	-.005	.004	.059	.001	.004	.071	.020	.005
	E_L	.058	-.004	.003	.055	-.002	.003	.061	-.005	.004	.060	.004	.004	.075	.001	.006
5000	E_M	.043	.002	.002	.033	-.003	.001	.035	-.002	.001	.040	-.006	.002	.043	.009	.002
	E_L	.043	.002	.002	.033	-.003	.001	.035	-.002	.001	.039	-.003	.002	.050	-.010	.002

The results of $\gamma_{8,2}$ in Table 3.12 are not influenced by the additional estimation of $\gamma_{8,1}$ and $\Omega_8^{(2,1)}$ in MINoLEM and `ltm`. Bias, variance, and RMSE of $\gamma_{8,1}$ are even lower compared to the other loadings in MINoLEM. In `ltm`, however, they are bigger. MINoLEM seems to estimate the zero-loading more accurately than `ltm`.

The results for the zero-interaction $\Omega_8^{(2,1)}$ are very good for both implementations. The values are lower than for the other two non-zero interaction terms, which are not affected. The biases, variances, and RMSEs decrease

Table 3.12: RMSE, bias, and variance of the coefficients of the latent variables of simulation for single-level model and correlation between the latent variables of 0. The loading $\gamma_{8,1}$ of the reading ability and the loading $\Omega_8^{(2,1)}$ of an interaction effect for item 8 are simulated to be 0 but estimated nevertheless.

N_C		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$			$\gamma_{8,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.264	.068	.065	.185	.008	.034	.179	.014	.032	.320	.023	.102	.182	-.006	.033
	E_L	.366	.076	.128	.185	.002	.034	.179	.005	.032	.375	.045	.139	.233	.143	.034
1000	E_M	.181	.026	.032	.137	.009	.019	.119	.006	.014	.200	.023	.040	.124	.003	.015
	E_L	.186	.021	.034	.136	.003	.019	.119	-.001	.014	.235	.032	.054	.157	.097	.015
2000	E_M	.140	.032	.019	.090	.006	.008	.086	.009	.007	.149	.020	.022	.089	-.006	.008
	E_L	.145	.027	.020	.088	.002	.008	.084	.003	.007	.166	.024	.027	.112	.069	.008
5000	E_M	.074	.007	.005	.058	.007	.003	.059	.014	.003	.089	.007	.008	.055	-.008	.003
	E_L	.074	-.000	.006	.057	.004	.003	.056	.010	.003	.095	.007	.009	.070	.044	.003

N_C		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$			$\gamma_{7,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.204	.022	.041	.351	-.022	.123	.258	.024	.010	.298	.050	.086	.234	.009	.055
	E_L	.211	.029	.044	.553	.095	.297	.278	.032	.010	.327	.053	.104	.243	.001	.059
1000	E_M	.146	.014	.021	.226	-.029	.050	.172	.010	.010	.169	.012	.028	.160	.002	.025
	E_L	.147	.017	.021	.308	.053	.092	.177	.011	.010	.172	.008	.030	.160	-.005	.026
2000	E_M	.088	.000	.008	.176	-.065	.027	.117	.007	.010	.120	.017	.014	.118	.017	.014
	E_L	.089	.003	.008	.190	-.006	.036	.118	.006	.010	.121	.013	.015	.116	.012	.013
5000	E_M	.058	-.003	.003	.115	-.036	.012	.078	-.003	.010	.079	.006	.006	.069	.005	.005
	E_L	.058	-.001	.003	.134	.023	.018	.079	-.003	.010	.080	.004	.006	.067	.000	.005

N_C		$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.185	.016	.034	.275	.088	.068	.243	-.002	.059
	E_L	.191	.017	.036	.301	.097	.081	.362	.068	.126
1000	E_M	.128	.011	.016	.179	.045	.030	.181	-.006	.033
	E_L	.129	.012	.017	.184	.047	.031	.236	.041	.054
2000	E_M	.091	.005	.008	.120	.019	.014	.114	-.030	.012
	E_L	.091	.005	.008	.122	.021	.014	.121	.001	.015
5000	E_M	.055	.003	.003	.066	.005	.004	.076	-.016	.006
	E_L	.055	.003	.003	.066	.005	.004	.086	.014	.007

with increased numbers of individuals.

Overall, both implementations can correctly distinguish between items that are influenced by both latent variables and their interaction and items that are only affected by one latent variable.

Table 3.13: RMSE, bias, and variance of the coefficients of the interaction between the latent variables, the correlation between the latent variables, and the variance of the random intercept of simulation for single-level model and correlation between the latent variables of 0. The loading $\gamma_{8,1}$ of the reading ability and the loading $\Omega_8^{(2,1)}$ of an interaction effect for item 8 are simulated to be 0 but estimated nevertheless.

N_C		$\Omega_8^{(2,1)}$			$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.292	.023	.085	.389	.070	.147	.440	-.031	.192
	E_L	.309	.011	.095	.438	.234	.137	.833	.158	.670
1000	E_M	.187	-.017	.035	.239	.041	.056	.343	-.056	.115
	E_L	.190	-.006	.036	.218	.106	.036	.479	.079	.223
2000	E_M	.126	.007	.016	.170	.022	.028	.229	-.070	.048
	E_L	.128	.012	.016	.140	.052	.017	.278	.026	.077
5000	E_M	.076	-.006	.006	.105	.029	.010	.159	-.048	.023
	E_L	.076	-.001	.006	.080	.013	.006	.198	.042	.037

3.2.4 Single-Level Data with Multilevel Model Specification

Before presenting the simulations for the full multilevel model, it is shown that MINoLEM is able to identify a missing hierarchical structure and still estimate the single-level model properly. This in turn gives the estimation of non-zero variances of a random intercept more weight in applications.

A model with hierarchical structure was estimated, while the data were simulated as a single-level model. The datasets were drawn with the same parameters as in Equations (3.1) and (3.2) with correlation $\rho = 0.3$ between the latent variables. The sizes of the datasets were set to $N = 2500, 5000, 7500$, so that the model is estimated assuming $N_C = 50$ clusters with $N_S = 50, 100, 150$ individuals each.

The RMSEs, biases, and variances of the item parameters and of the correlation between the latent variables in Tables 3.14, 3.15, and 3.16 are very similar to the results, in which no multilevel structure was assumed (e.g., in Tables B.3 and B.4 in the appendix and Table 3.10). The very low bias,

Table 3.14: RMSE, bias, and variance of the difficulties of a simulation, in which a hierarchical structure was assumed while none was present. The correlation between the latent variables was set to 0.3. The interaction terms have medium values.

N	δ_1			δ_2			δ_3			δ_4			δ_5		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
N=50·50	.058	.009	.003	.052	.003	.003	.044	.001	.002	.060	.009	.004	.065	-.002	.004
N=100·50	.042	.003	.002	.033	.000	.001	.030	-.003	.001	.038	.004	.001	.045	.008	.002
N=150·50	.036	.000	.001	.029	.002	.001	.025	-.002	.001	.034	.002	.001	.038	.003	.001

N	δ_6			δ_7			δ_8			δ_9			δ_{10}		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
N=50·50	.058	-.003	.003	.046	-.002	.002	.054	-.004	.003	.055	-.001	.003	.072	-.007	.005
N=100·50	.042	-.006	.002	.035	-.002	.001	.035	.002	.001	.039	-.001	.002	.050	-.011	.002
N=150·50	.036	-.005	.001	.029	.005	.001	.030	.002	.001	.033	-.004	.001	.042	-.009	.002

Table 3.15: RMSE, bias, and variance of the coefficients of the latent variables of a simulation, in which a hierarchical structure was assumed while none was present. The correlation between the latent variables was set to 0.3. The interaction terms have medium values.

N	$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
N=50·50	.106	.010	.011	.078	.002	.006	.086	.005	.007	.134	.008	.018
N=100·50	.077	.005	.006	.056	.001	.003	.054	.008	.003	.087	.011	.007
N=150·50	.063	-.000	.004	.044	-.001	.002	.041	.002	.002	.080	.021	.006

N	$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
N=50·50	.102	.030	.009	.151	-.014	.023	.105	.004	.011	.107	.010	.001
N=100·50	.064	.015	.004	.120	-.038	.013	.073	.017	.005	.079	.008	.001
N=150·50	.058	.014	.003	.099	-.033	.009	.059	.010	.003	.067	.014	.001

N	$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
N=50·50	.092	-.008	.008	.075	.018	.005	.117	.001	.014	.124	-.014	.015
N=100·50	.059	.002	.003	.047	-.001	.002	.073	.002	.005	.081	-.003	.007
N=150·50	.047	.003	.002	.038	-.003	.001	.063	.006	.004	.069	-.007	.005

variance, and RMSE for the variance σ^2 of the random intercept in Table 3.16 suggests that the alleged variance of the random intercept is correctly estimated close to 0 – consistently for every condition. MINoLEM recognizes the missing multilevel structure, while estimating all the other parameters correctly as well.

Table 3.16: RMSE, bias, and variance of the difficulties of a simulation, in which a hierarchical structure was assumed while none was present. The correlation between the latent variables was set to 0.3. The interaction terms have medium values.

N	$\Omega_9^{(2,1)}$			$\Omega_9^{(2,1)}$			ρ			σ^2		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
N=50·50	.178	-.002	.032	.230	-.043	.051	.053	-.019	.002	.001	.000	.000
N=100·50	.121	.009	.015	.167	-.060	.024	.042	-.016	.001	.000	.000	.000
N=150·50	.096	.005	.009	.148	-.053	.019	.036	-.015	.001	.000	.000	.000

3.3 Simulation Study – Multilevel

In the final simulation, the complete model with the consideration of hierarchical data as in Equation (2.1) is examined. To investigate the consistency of the estimation, the results are presented in two parts. First, the number of individuals per cluster N_S is fixed, which allows to observe the influence of an increasing number of clusters. Second, the number of clusters N_C is fixed, which allows to observe the influence of an increasing number of individuals per cluster. The results of MINoLEM (E_M) are compared to the estimates of `mirt` (E_m), a package in R (Chalmers, 2012). The interaction is only estimated by MINoLEM, but both account for the hierarchical structure.

The parameters were equal in both simulations. The difficulties for the 10 items were, once again, set to

$$\boldsymbol{\delta} = \begin{pmatrix} 1 & -1.2 & -0.2 & 0.6 & 1.2 & -0.6 & 0.2 & -1 & 0 & -0.4 \end{pmatrix}^t ,$$

while a two-dimensional latent variable with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_2})$ was assumed. Each latent variable had four items that depended only on them – the first four on the first latent variable dimension and the following four on the second dimension. The last two items were affected by both latent variable dimensions

$$\boldsymbol{\gamma} = \begin{pmatrix} 1 & 0.5 & 0.55 & 1.2 & 0 & 0 & 0 & 0 & 0.45 & 1.1 \\ 0 & 0 & 0 & 0 & 1 & 1.15 & 0.95 & 0.6 & 1.05 & 0.65 \end{pmatrix}^t ,$$

as well as by the interaction of the latent variables

$$\boldsymbol{\Omega}_9 = \begin{pmatrix} 0 & 0 \\ 0.45 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega}_{10} = \begin{pmatrix} 0 & 0 \\ 1.1 & 0 \end{pmatrix} .$$

The variance of the random intercept was set to $\sigma^2 = 0.125$, and the correlation between the latent variables to $\rho = 0.3$.

3.3.1 Fixed Size of Clusters

A multilevel model with varying numbers of clusters $N_C = 50, 100, 200$ and a fixed cluster size to $N_S = 100$ was simulated. The minimum number of clusters is in accordance with McNeish and Stapleton (2016), who suggest at least 50 clusters to estimate the Level 2 variance (for estimation with FIML). They also suggest a cluster size of at least 10 for Level 1 fixed effects, which is later increased to $N_S = 50$, since nonlinear effects usually need more resources to be estimated accurately.

Difficulties of the Items

The biases of the difficulties in Table 3.17 are very small for the estimates of MINoLEM (E_M) for all numbers of clusters. The estimates of `mirt` (E_m) show a slightly higher bias for $N_C = 50$ and $N_C = 100$, but the bias decreases with rising numbers of clusters and is in the same order of magnitude as those of MINoLEM when $N_C = 200$. The only exception is the bias of the difficulty δ_{10} of item ten. MINoLEM shows slightly higher values than for the other item difficulties. The biases of `mirt` increase with higher numbers of

Table 3.17: RMSE, bias, and variance of difficulties of simulation for multilevel model and correlation between the latent variables of 0.3. The number of individuals per cluster is fixed to $N_S = 100$. N_C = Number of clusters. E_M = Estimation with MINoLEM. E_m = Estimation with `mirt`.

N_C		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.062	.003	.004	.059	-.001	.003	.056	-.001	.003	.065	.002	.004	.062	.006	.004
	E_m	.076	.015	.006	.071	.014	.005	.068	.013	.004	.079	.015	.006	.076	.019	.005
100	E_M	.047	.007	.002	.041	-.002	.002	.041	.002	.002	.045	.005	.002	.042	.003	.002
	E_m	.054	.014	.003	.050	.012	.002	.050	.014	.002	.053	.012	.003	.049	.012	.002
200	E_M	.032	-.002	.001	.031	-.005	.001	.029	-.001	.001	.032	-.002	.001	.033	.001	.001
	E_m	.037	-.003	.001	.034	.001	.001	.032	.002	.001	.037	-.003	.001	.038	.001	.001

N_C		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.061	-.001	.004	.060	-.001	.004	.061	-.004	.004	.056	-.002	.003	.067	-.010	.004
	E_m	.075	.015	.005	.074	.014	.005	.070	.011	.005	.068	.006	.005	.079	-.030	.005
100	E_M	.044	-.001	.002	.041	.002	.002	.046	.003	.002	.043	.002	.002	.046	-.010	.002
	E_m	.052	.012	.003	.050	.012	.002	.055	.015	.003	.049	.004	.002	.066	-.035	.003
200	E_M	.032	-.005	.001	.030	-.001	.001	.030	-.005	.001	.031	-.004	.001	.040	-.015	.001
	E_m	.035	-.001	.001	.034	.000	.001	.033	-.002	.001	.037	-.010	.001	.077	-.049	.004

clusters and are significantly higher than those of MINoLEM. The package `mirt`, however, estimates a misspecified model, since the interactions are not taken into account. That explains the higher bias of `mirt`. Surprisingly, this effect is not as big for δ_9, where the interaction is also not estimated in `mirt`. Both methods show clear signs of consistency: Whenever the number of clusters is increased, the bias is decreased.

The variances and the RMSEs of the estimates decrease for all difficulties and for both methods. The only exceptions are again the values of `mirt` for δ_{10}, which are slightly higher due to the misspecification.

Loadings of the Latent Variables

The biases of the loadings of the latent variables in Table 3.18 are very small for both methods and all conditions. The order of magnitude of the biases are similar to the difficulties' biases. More than 10 times larger values can, once again, be observed in `mirt` for the loadings of the latent variables for item ten ($\gamma_{10,1}$ and $\gamma_{10,2}$). Since the biases for the cross-loadings of the first latent dimension do not increase in `mirt`, the misspecification seems to 'push' all bias to one item only. The biases of MINoLEM for $\gamma_{10,1}$ and $\gamma_{10,2}$ are also slightly higher than the biases for the other latent variable coefficients, which can be explained by the additional dependence of the item on the interaction.

Table 3.18: RMSE, bias, and variance of the coefficients of the latent variables of simulation for multilevel model and correlation between the latent variables of 0.3. The number of individuals per cluster is fixed to $N_S = 100$. N_C = Number of clusters. E_M = Estimation with MINoLEM. E_m = Estimation with `mirt`.

N_C		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.079	.007	.006	.050	.002	.002	.051	.011	.003	.085	.009	.007
	E_m	.084	.001	.007	.055	.002	.003	.055	.006	.003	.095	.001	.009
100	E_M	.052	.004	.003	.038	.003	.001	.038	.001	.001	.064	.012	.004
	E_m	.057	-.012	.003	.041	-.006	.002	.042	-.007	.002	.070	-.015	.005
200	E_M	.039	.011	.001	.029	.007	.001	.026	.005	.001	.047	.014	.002
	E_m	.038	-.004	.001	.028	-.002	.001	.026	-.004	.001	.049	-.010	.002

N_C		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.069	.024	.004	.124	-.053	.013	.075	.007	.006	.075	.016	.005
	E_m	.066	.017	.004	.375	-.259	.073	.076	.005	.006	.079	.014	.006
100	E_M	.054	.017	.003	.097	-.044	.007	.049	.005	.002	.053	.014	.003
	E_m	.051	.001	.003	.378	-.264	.073	.051	.003	.003	.053	.003	.003
200	E_M	.040	.015	.001	.084	-.045	.005	.033	.004	.001	.038	.006	.001
	E_m	.037	.009	.001	.383	-.270	.074	.032	-.001	.001	.039	.000	.001

N_C		$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.063	.009	.004	.055	.008	.003	.078	.010	.006	.085	-.025	.007
	E_m	.066	.006	.004	.056	.008	.003	.084	.008	.007	.347	-.240	.063
100	E_M	.047	.005	.002	.037	.002	.001	.054	.005	.003	.066	-.023	.004
	E_m	.048	.000	.002	.039	.000	.002	.059	.007	.003	.347	-.243	.062
200	E_M	.034	.005	.001	.024	.002	.001	.038	.006	.001	.047	-.019	.002
	E_m	.035	.000	.001	.025	-.001	.001	.044	.007	.002	.333	-.234	.056

For both methods, the biases are reducing with increasing numbers of clusters, although not as clearly as for the difficulties. Signs of consistency are visible but ambiguous for some coefficients.

The RMSEs and the variances, however, decline with rising numbers of clusters for both methods (except in `mirt` for item ten).

Overall, both methods show very similar values for the biases, variances, and RMSEs of the difficulties and the loadings, if the increased values, due to the misspecification in `mirt`, are not taken into account.

Interaction Coefficients, Correlation, and Random Intercept

The biases of the first interaction coefficient $\Omega_9^{(2,1)}$ in Table 3.19 are very low and of the order of magnitude as the biases of the loadings of the main effects. The biases of the second interaction coefficient $\Omega_{10}^{(2,1)}$ are higher, due to the higher true value, as was discussed in Section 3.2.1. Both biases stay relatively constant with rising numbers of clusters, which might indicate that the number of clusters has no or only little influence on the estimates of the interaction coefficients, contrary to the estimates of the difficulties and the loadings.

Table 3.19: RMSE, bias, and variance of the coefficients of the interaction between the latent variables, the correlation between the latent variables, and the variance of the random intercept of simulation for multilevel model and correlation between the latent variables of 0.3. The number of individuals per cluster is fixed to $N_S = 100$. N_C = Number of clusters. E_M = Estimation with MINoLEM. E_m = Estimation with `mirt`.

N_C		$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$			ρ			σ^2		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.118	.001	.014	.201	-.103	.030	.048	-.026	.002	.027	-.012	.001
	E_m							.039	.001	.002	.037	.002	.001
100	E_M	.084	.004	.007	.168	-.092	.020	.040	-.022	.001	.022	-.012	.000
	E_m							.030	.009	.001	.022	-.002	.000
200	E_M	.055	.004	.003	.150	-.091	.014	.034	-.020	.001	.018	-.010	.000
	E_m							.020	-.002	.000	.014	-.002	.000

However, both the variances and the RMSEs decrease, which shows that the estimation becomes more accurate with an increase of the number of clusters.

The correlation between the latent variables has a slightly higher bias in MINoLEM than in `mirt`, but it decreases with rising numbers of clusters. Both methods have equally low variances and a declining RMSEs.

The variance of the random intercept is, once again, more accurately estimated by `mirt` regarding the bias. The seemingly contradicting values of the RMSE (lower for MINoLEM than for `mirt`, although the bias and the variance are lower in `mirt`) are caused by rounding – the variances of MINoLEM for $N_C = 50$ and $N_C = 100$ are lower than the variances of `mirt`.

In summary, both methods show a very good performance. The package `mirt` seems to cope with the misspecifation, while MINoLEM seems to accurately estimate the interaction coefficients, as well as the multilevel structure in form of a random intercept. No significant difference between the two compared methods can be observed regarding the estimation of the difficulties, the loadings of the latent variables, the correlation between the latent variables, and the random intercept.

3.3.2 Fixed Number of Clusters

The total number of individuals can be increased by either adding more clusters, or by adding more individuals to each cluster. To investigate the second possibility, the number of clusters is fixed to $N_C = 100$ and the number of individuals per cluster varies between $N_S = 50, 100, 150$. The true values are chosen as before.

Difficulties of the Items

MINoLEM has almost the same results as for fixed N_S. The biases decrease with rising numbers of individuals per cluster and show clear signs of consistency. The biases of `mirt` decrease as well, however not as much as before and they are consistently higher than the biases of MINoLEM. The biases of the difficulty δ_{10} of item ten are, once again, slightly higher for `mirt`, due to the misspecification.

The variances and the RMSEs of both methods are small and decrease, as they did for a fixed number of clusters.

Table 3.20: RMSE, bias, and variance of difficulties of simulation for multi-level model and correlation between the latent variables of 0.3. The number of cluster is fixed to $N_C = 100$. N_C = Number of clusters. E_M = Estimation with MINoLEM. E_m = Estimation with `mirt`.

N_S		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.055	.008	.003	.049	-.001	.002	.045	-.001	.002	.053	-.002	.003	.061	-.008	.004
	E_m	.067	.016	.004	.059	.016	.003	.058	.014	.003	.061	.007	.004	.066	.006	.004
100	E_M	.047	.007	.002	.041	-.002	.002	.041	.002	.002	.045	.005	.002	.042	.003	.002
	E_m	.054	.014	.003	.050	.012	.002	.050	.014	.002	.053	.012	.003	.049	.012	.002
150	E_M	.038	.003	.001	.036	-.001	.001	.036	-.002	.001	.038	-.000	.001	.042	.002	.002
	E_m	.046	.009	.002	.047	.014	.002	.046	.011	.002	.047	.007	.002	.051	.011	.002

N_S		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.053	-.005	.003	.047	-.001	.002	.051	.004	.003	.055	-.004	.003	.067	-.016	.004
	E_m	.063	.011	.004	.057	.013	.003	.063	.020	.004	.059	.001	.003	.078	-.039	.005
100	E_M	.044	-.001	.002	.041	.002	.002	.046	.003	.002	.043	.002	.002	.046	-.010	.002
	E_m	.052	.012	.003	.050	.012	.002	.055	.015	.003	.049	.004	.002	.066	-.035	.003
150	E_M	.038	-.003	.001	.039	-.004	.001	.038	-.003	.001	.040	-.005	.002	.042	-.012	.002
	E_m	.047	.012	.002	.047	.008	.002	.049	.011	.002	.047	-.000	.002	.066	-.036	.003

Loadings of the Latent Variables

All three measured features of the estimates, RMSE, bias, and variance still basically do not change for both methods. For the estimation of the latent variable coefficients, it does not seem to matter whether the total number of individuals is increased by adding more clusters or by adding more individuals per cluster.

Table 3.21: RMSE, bias, and variance of the coefficients of the latent variables of simulation for multilevel model and correlation between the latent variables of 0.3. The number of cluster is fixed to $N_C = 100$. N_C = Number of clusters. E_M = Estimation with MINoLEM. E_m = Estimation with `mirt`.

N_S		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.079	.018	.006	.054	-.001	.003	.054	.005	.003	.084	.017	.007
	E_m	.079	-.000	.006	.057	-.011	.003	.057	-.006	.003	.085	-.003	.007
100	E_M	.052	.004	.003	.038	.003	.001	.038	.001	.001	.064	.012	.004
	E_m	.057	-.012	.003	.041	-.006	.002	.042	-.007	.002	.070	-.015	.005
150	E_M	.047	.011	.002	.035	.007	.001	.032	.008	.001	.052	.012	.003
	E_m	.046	-.006	.002	.036	-.006	.001	.032	-.003	.001	.055	-.009	.003

N_S		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.070	.013	.005	.123	-.035	.014	.073	-.002	.005	.083	.016	.007
	E_m	.070	.001	.005	.396	-.274	.082	.077	-.001	.006	.089	.013	.008
100	E_M	.054	.017	.003	.097	-.044	.007	.049	.005	.002	.053	.014	.003
	E_m	.051	.001	.003	.378	-.264	.073	.051	.003	.003	.053	.003	.003
150	E_M	.049	.019	.002	.105	-.059	.007	.043	.008	.002	.046	.008	.002
	E_m	.044	.007	.002	.394	-.277	.079	.044	-.003	.002	.048	-.005	.002

N_S		$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.064	.012	.004	.056	-.000	.003	.077	.003	.006	.083	-.008	.007
	E_m	.067	.010	.004	.057	-.001	.003	.083	.010	.007	.333	-.231	.058
100	E_M	.047	.005	.002	.037	.002	.001	.054	.005	.003	.066	-.023	.004
	E_m	.048	.000	.002	.039	.000	.002	.059	.007	.003	.347	-.243	.062
150	E_M	.040	.009	.002	.034	.010	.001	.044	.002	.002	.051	-.019	.002
	E_m	.039	-.003	.002	.032	.003	.001	.047	-.003	.002	.332	-.234	.056

Interaction Coefficients, Correlation, and Random Intercept

The coefficients of the latent variable interaction seem to be estimated slightly more accurately when the number of individuals per cluster is fixed with an increasing number of clusters. The biases and the variances are marginally higher in this condition, except for the values of $\Omega_{10}^{(2,1)}$ when $N_S = 50$, which are smaller.

The correlation between the latent variables is estimated equally well by both methods as in the previous section.

When the number of individuals is increased, the estimation of the variance of the random intercept, however, produces a slightly rising bias for MINoLEM. The package `mirt` produces the same results as with varying numbers of clusters.

Table 3.22: RMSE, bias, and variance of the coefficients of the interaction between the latent variables, the correlation between the latent variables, and the variance of the random intercept of simulation for multilevel model and correlation between the latent variables of 0.3. The number of clusters is fixed to $N_C = 100$. N_C = Number of clusters. E_M = Estimation with MINoLEM. E_m = Estimation with `mirt`.

N_S		$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$			ρ			σ^2		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	E_M	.129	.008	.017	.178	-.072	.027	.043	-.017	.002	.019	-.002	.000
	E_m							.043	.005	.002	.023	-.003	.001
100	E_M	.084	.004	.007	.168	-.092	.020	.040	-.022	.001	.022	-.012	.000
	E_m							.030	.009	.001	.022	-.002	.000
150	E_M	.069	.010	.005	.174	-.106	.019	.039	-.023	.001	.030	-.020	.001
	E_m							.023	-.001	.001	.021	-.001	.000

Overall, most results show little to no difference, if either the number of clusters or the number of individuals per cluster is fixed, while the other is increased.

The results demonstrate that MINoLEM is capable of estimating a multilevel IRT model with interaction effects of the latent variables. The comparison to the R packages `ltm` and `mirt` show that MINoLEM performs equally well as those established implementations. Furthermore, MINoLEM can help to extend the range of IRT models that can be estimated.

Chapter 4

An Application to PISA Data

In the previous chapter, extensive simulations showed that the estimation with MINoLEM produces results that are comparable to the established packages mirt and ltm. MINoLEM extends them by adding interaction effects and hierarchical structures, respectively, which are estimated with low bias, variance, and RMSE. In this chapter, PISA data are evaluated with MINoLEM to give a proof of concept and show its applicability to real data.

4.1 Introduction to PISA

According to the website of the OECD, PISA is described as "the OECD's Programme for International Student Assessment. PISA measures 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges" (OECD, 2021).

The first PISA survey was done in 2000 and was repeated every three years since then in each OECD country and in a rising number of other countries as well. One goal of the study is to provide comparable results of children's abilities between countries. OECD Secretary-General Angel Gurría elaborated that

> *"PISA is not only the world's most comprehensive and reliable indicator of students' capabilities, it is also a powerful tool that countries and economies can use to fine-tune their education policies" (Schleicher, 2018).*

In many countries, the results of PISA are discussed in detail in the media, and changes in the educational system are demanded. Especially in Germany, the media echo was immense after the first three surveys (Popp, 2010). As a response, the Kultusministerkonferenz (conference of ministers of education) decided on German wide learning standards and founded the Institut zur Qualitätsentwicklung im Bildungswesen (Institute for Educational Quality Improvement - IQB) in 2004 (IQB, 2021).

However, since the beginning of PISA, many have criticized the methods, with which the results are achieved. Roughly speaking, two estimation procedures are performed. First, IRT is applied to evaluate the items. For each of the three fields, reading, mathematics, and science knowledge, an underlying factor is assumed that follows a normal distribution and that affects the probability of answering an item correctly. In the beginning, the relationship between the factors and the items was estimated with a Rasch model. Critics raised concern that more complex models are needed and it could be proven that the results change significantly (e.g., Kreiner and Christensen (2014)). Today, the relationships are estimated with a 2PL model.

Second, plausible values are drawn from a posterior probability of the abilities of the students that resulted from the IRT evaluation. Those plausible values represent random student scores. The plausible values are then used to rank the participating countries according to the abilities of their students. Much more detailed descriptions can be found in the technical report from the OECD for each PISA study (e.g., of the OECD (2017)).

Among other aspects, the inspection PISA's items does not account for the hierarchical structure of different educational systems around the world.

Furthermore, no cross-loadings or interactions between the latent abilities are allowed. Therefore, the MINoLEM model will be applied to PISA data to explore, whether those extensions could improve the results of the PISA surveys.

4.2 Estimating parameters in OECD Countries

Here, the data from the survey 2006 were chosen, in which 22 reading items and 44 math items were administered. The study was conducted in all 30 OECD countries and in additional 27 other countries, one of which (Lichtenstein) was excluded from the analyses, because it had too few participants. The missing values were imputed with the `amelia` package (Honaker, King, & Blackwell, 2011) using an EM algorithm.

Two different sets of countries were analyzed – all 56 countries together (next section) and only the OECD countries (this section) – to see how the variance of the random intercept might change. To give every country the same weight, 100 individuals from each country were sampled.

The reading items are assumed to be influenced only by reading ability of the students. The performance in the math items, however, might also depend on reading ability of the students. Therefore, 34 of the 44 math items were set to be influenced by an interaction of reading and math ability and by both main effects. Ten math items were chosen, which were defined to only depend on math ability to build a stable factor. Unfortunately, the item formulations are not publicly available, so that it is not possible to choose those items that might depend the least on reading ability. Furthermore, it is possible that the influences change between different languages so that instead the first ten items are chosen to build the factor for math ability.

The estimates of all 66 difficulties in Table 4.1 range from -2.42 to 2.38 and approximately follow a bell shaped distribution (see Figure 4.1) with a slight shift to the left of 0, so that more items could be labeled easy than difficult. This indicates a reasonably well estimation of the difficulties.

Table 4.1: Estimated difficulties of the reading and math items of the PISA data (2006) in the OECD countries.

δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}	δ_{14}	δ_{15}	δ_{16}	δ_{17}
-1.28	.10	-.33	-.95	-1.60	.67	.22	-1.44	-1.24	.52	-.51	-.21	-.80	-1.25	.29	-.37	-.34

δ_{18}	δ_{19}	δ_{20}	δ_{21}	δ_{22}	δ_{23}	δ_{24}	δ_{25}	δ_{26}	δ_{27}	δ_{28}	δ_{29}	δ_{30}	δ_{31}	δ_{32}	δ_{33}	δ_{34}
-1.31	-.52	-.10	-.20	-.85	-1.04	.21	-.56	-.27	.36	-.24	-2.42	-1.17	.89	-.38	.97	1.54

δ_{35}	δ_{36}	δ_{37}	δ_{38}	δ_{39}	δ_{40}	δ_{41}	δ_{42}	δ_{43}	δ_{44}	δ_{45}	δ_{46}	δ_{47}	δ_{48}	δ_{49}	δ_{50}	δ_{51}
.12	-.11	.22	-.03	-.49	1.50	.72	-1.17	.37	-.79	2.38	-.74	.98	-.91	-.08	-.60	-.55

δ_{52}	δ_{53}	δ_{54}	δ_{55}	δ_{56}	δ_{57}	δ_{58}	δ_{59}	δ_{60}	δ_{61}	δ_{62}	δ_{63}	δ_{64}	δ_{65}	δ_{66}
.13	.19	.10	-.29	.18	.55	.71	-1.64	.91	-.41	-.66	.52	-.18	.83	.73

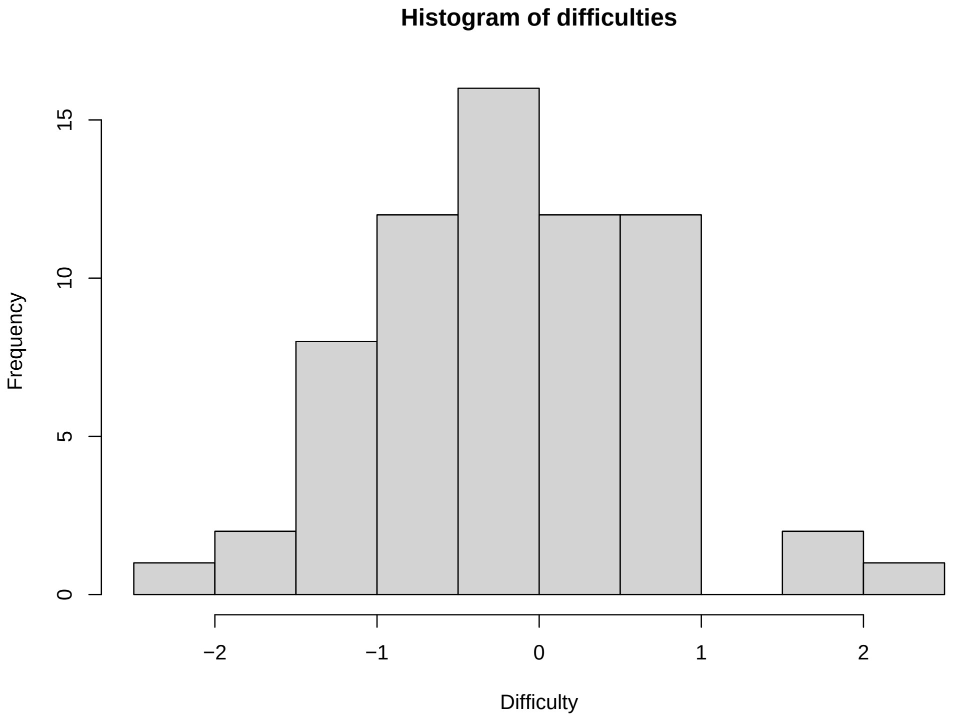

Figure 4.1: Histogram of the estimated difficulties for the OECD PISA data for all 66 items.

The estimates of the loadings of reading and math ability in Tables 4.2 and 4.3 are predominantly positive. Only a few items have loadings close to zero that should be investigated further.

Table 4.2: Estimated coefficients of the reading items of the PISA data (2006) in the OECD countries. The coefficients $\gamma_{33,1}$ to $\gamma_{66,1}$ are cross-loadings.

$\gamma_{1,1}$	$\gamma_{2,1}$	$\gamma_{3,1}$	$\gamma_{4,1}$	$\gamma_{5,1}$	$\gamma_{6,1}$	$\gamma_{7,1}$	$\gamma_{8,1}$	$\gamma_{9,1}$	$\gamma_{10,1}$	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{13,1}$	$\gamma_{14,1}$
.64	.93	1.07	1.07	.64	.78	.91	.72	.80	.44	.94	1.19	1.15	.97

$\gamma_{15,1}$	$\gamma_{16,1}$	$\gamma_{17,1}$	$\gamma_{18,1}$	$\gamma_{19,1}$	$\gamma_{20,1}$	$\gamma_{21,1}$	$\gamma_{22,1}$	$\gamma_{33,1}$	$\gamma_{34,1}$	$\gamma_{35,1}$	$\gamma_{36,1}$	$\gamma_{37,1}$	$\gamma_{38,1}$
1.10	.83	.77	1.00	.70	.54	.98	1.17	-.25	-.14	.19	.52	.08	.28

$\gamma_{39,1}$	$\gamma_{40,1}$	$\gamma_{41,1}$	$\gamma_{42,1}$	$\gamma_{43,1}$	$\gamma_{44,1}$	$\gamma_{45,1}$	$\gamma_{46,1}$	$\gamma_{47,1}$	$\gamma_{48,1}$	$\gamma_{49,1}$	$\gamma_{50,1}$	$\gamma_{51,1}$	$\gamma_{52,1}$
.57	-.28	-.22	.04	.35	.64	-.25	.25	-.14	.15	.05	.11	.20	.04

$\gamma_{53,1}$	$\gamma_{54,1}$	$\gamma_{55,1}$	$\gamma_{56,1}$	$\gamma_{57,1}$	$\gamma_{58,1}$	$\gamma_{59,1}$	$\gamma_{60,1}$	$\gamma_{61,1}$	$\gamma_{62,1}$	$\gamma_{63,1}$	$\gamma_{64,1}$	$\gamma_{65,1}$	$\gamma_{66,1}$
.04	.25	.28	.21	.25	.02	.27	-.10	.58	.37	.35	.36	.06	.09

Table 4.3: Estimated coefficients of the math items of the PISA data (2006) in the OECD countries. The coefficients $\gamma_{33,2}$ to $\gamma_{66,2}$ are cross-loadings.

$\gamma_{23,2}$	$\gamma_{24,2}$	$\gamma_{25,2}$	$\gamma_{26,2}$	$\gamma_{27,2}$	$\gamma_{28,2}$	$\gamma_{29,2}$	$\gamma_{30,2}$	$\gamma_{31,2}$	$\gamma_{32,2}$	$\gamma_{33,2}$	$\gamma_{34,2}$	$\gamma_{35,2}$	$\gamma_{36,2}$	$\gamma_{37,2}$
.64	.85	.82	.67	.92	.67	.66	.76	1.19	.52	1.29	1.01	.69	.61	.64

$\gamma_{38,2}$	$\gamma_{39,2}$	$\gamma_{40,2}$	$\gamma_{41,2}$	$\gamma_{42,2}$	$\gamma_{43,2}$	$\gamma_{44,2}$	$\gamma_{45,2}$	$\gamma_{46,2}$	$\gamma_{47,2}$	$\gamma_{48,2}$	$\gamma_{49,2}$	$\gamma_{50,2}$	$\gamma_{51,2}$	$\gamma_{52,2}$
.50	.96	.84	.78	.40	.77	.44	.95	.69	1.15	.42	.92	.58	.59	.56

$\gamma_{53,2}$	$\gamma_{54,2}$	$\gamma_{55,2}$	$\gamma_{56,2}$	$\gamma_{57,2}$	$\gamma_{58,2}$	$\gamma_{59,2}$	$\gamma_{60,2}$	$\gamma_{61,2}$	$\gamma_{62,2}$	$\gamma_{63,2}$	$\gamma_{64,2}$	$\gamma_{65,2}$	$\gamma_{66,2}$
.58	.72	.39	.58	.75	.78	.13	1.24	.43	.54	.81	.50	.64	.59

Of most interest are the loadings of the interaction effects. Ten items have coefficients between $0.1 < |\Omega_i^{(2,1)}| < 0.2$. As can be seen in Figure 4.2, this adds a significant twist to the IRS. Therefore, the probabilities of answering those items correctly, seem to be influenced not only by students' math ability, but also by their reading ability and by the interaction of both factors.

The correlation between the two latent variables – math and reading ability – is estimated very high with $\rho = 0.72$. This might indicate that the measured concepts are not as distinct as assumed. The variance of the random intercept, however, is estimated low with $\sigma^2 = 0.03$, which implies that the OECD countries do not differ much between each other in their mean probability of answering an item correctly.

Table 4.4: Estimated coefficients of the interaction effects, the correlation between the latent variables, and the variance of the random intercept of the PISA data (2006) in the OECD countries.

$\Omega_{33}^{(2,1)}$	$\Omega_{34}^{(2,1)}$	$\Omega_{35}^{(2,1)}$	$\Omega_{36}^{(2,1)}$	$\Omega_{37}^{(2,1)}$	$\Omega_{38}^{(2,1)}$	$\Omega_{39}^{(2,1)}$	$\Omega_{40}^{(2,1)}$	$\Omega_{41}^{(2,1)}$	$\Omega_{42}^{(2,1)}$	$\Omega_{43}^{(2,1)}$	$\Omega_{44}^{(2,1)}$
.04	.20	-.04	.11	.10	-.05	.16	.10	.16	-.02	.00	-.13
$\Omega_{45}^{(2,1)}$	$\Omega_{46}^{(2,1)}$	$\Omega_{47}^{(2,1)}$	$\Omega_{48}^{(2,1)}$	$\Omega_{49}^{(2,1)}$	$\Omega_{50}^{(2,1)}$	$\Omega_{51}^{(2,1)}$	$\Omega_{52}^{(2,1)}$	$\Omega_{53}^{(2,1)}$	$\Omega_{54}^{(2,1)}$	$\Omega_{55}^{(2,1)}$	$\Omega_{56}^{(2,1)}$
.19	-.03	.02	-.05	-.04	-.07	.03	.10	.09	-.00	.04	.06
$\Omega_{57}^{(2,1)}$	$\Omega_{58}^{(2,1)}$	$\Omega_{59}^{(2,1)}$	$\Omega_{60}^{(2,1)}$	$\Omega_{61}^{(2,1)}$	$\Omega_{62}^{(2,1)}$	$\Omega_{63}^{(2,1)}$	$\Omega_{64}^{(2,1)}$	$\Omega_{65}^{(2,1)}$	$\Omega_{66}^{(2,1)}$	ρ	σ^2
.01	.09	.01	.17	.05	-.01	.01	.00	.03	.04	.72	.03

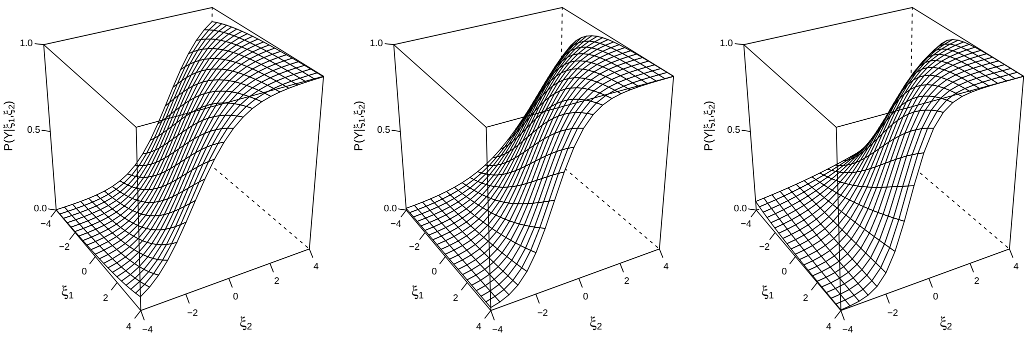

Figure 4.2: Item response surfaces of 2-dimensional IRT model item with interaction effects $\Omega = 0$, $\Omega = 0.1$, and $\Omega = 0.2$, respectively.

Standard errors are not presented for this estimation, due to the prohibitive amount of computing time needed by the current version of MINoLEM to estimate a model with 202 parameters. To provide a model with standard errors, the number of items was reduced to twelve. Four items are assumed to be influenced only by either reading ability or math ability, respectively. Four items depend on both main effects and on their interaction. Those items with the most matching values in the previous estimation were chosen, so that items $12, 13, 15$, and 22, define reading ability, items $33, 47, 57$, and 63 define math ability, and items $34, 44, 45$, and 60 will be assumed to depend on both latent variables and their interaction.

Table 4.5: Estimated difficulties of the reading and math items of the reduced PISA data (2006) in the OECD countries. The standard errors are given in brackets.

δ_{12}	δ_{13}	δ_{15}	δ_{22}	δ_{33}	δ_{47}	δ_{57}	δ_{63}	δ_{34}	δ_{44}	δ_{45}	δ_{60}
-.302(.052)	-.875(.065)	.286(.042)	-.790(.047)	1.061(.061)	.984(.052)	.548(.047)	.465(.044)	1.895(.138)	-.715(.059)	2.195(.093)	.747(.059)

The difficulties of the reduced dataset in Table 4.5 approximately match those of the complete data. The estimated loadings in Table 4.6 are also approximately the same, except for $\gamma_{34,1}$ and $\gamma_{34,2}$.

Table 4.6: Estimated loadings of the reading and math items of the reduced PISA data (2006) in the OECD countries. The standard errors are given in brackets.

$\gamma_{12,1}$	$\gamma_{13,1}$	$\gamma_{15,1}$	$\gamma_{22,1}$	$\gamma_{34,1}$	$\gamma_{44,1}$	$\gamma_{45,1}$	$\gamma_{60,1}$	$\gamma_{33,2}$	$\gamma_{47,2}$	$\gamma_{57,2}$	$\gamma_{63,2}$	$\gamma_{34,2}$	$\gamma_{44,2}$	$\gamma_{45,2}$	$\gamma_{60,2}$
1.610(.127)	1.710(.149)	.835(.077)	.855(.071)	-.912(.252)	.544(.100)	-.247(.140)	.064(.110)	1.439(.096)	1.031(.072)	.800(.063)	.934(.069)	2.583(.283)	.425(.096)	.907(.137)	.923(.101)

The correlation of the latent variables and the variance of the random intercept in Table 4.7 change only slightly. The standard errors suggest that the 95% confidence intervals of all those values do not include 0 and are therefore significant (on a $\alpha = 0.05$ error level), except for the two values $\gamma_{45,1}$ and $\gamma_{60,1}$, which are too small and not significant.

The different operationalizations of the latent variables, however, affect the coefficients of the interaction effects. None of them are close to the estimated

Table 4.7: Estimated loadings of the interaction effects, the correlation between the latent variables, and the variance of the random intercept of the reduced PISA data (2006) in the OECD countries. The standard errors are given in brackets.

$\Omega_{9}^{(24,1)}$	$\Omega_{10}^{(44,1)}$	$\Omega_{45}^{(2,1)}$	$\Omega_{60}^{(2,1)}$	ρ	σ^2
-.122(.188)	.003(.086)	.050(.101)	.079(.087)	.676(.039)	.050(.011)

values of the complete dataset. Since all four values are very small, none of them are significant. Nevertheless, the application demonstrates that further investigation of the PISA data is necessary, and that MINoLEM could be one way to reevaluate the PISA analysis.

4.3 Estimating parameters in all Countries

In the previous section, the PISA data was examined for the OECD countries. In this simulation, the PISA data are examined for all participating countries, including OECD, to show that the hierarchical aspect of MINoLEM can be of importance in different circumstances.

Table 4.8: Estimated difficulties of the reading and math items of the PISA data (2006) for all countries.

δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}	δ_{11}	δ_{12}	δ_{13}	δ_{14}	δ_{15}	δ_{16}	δ_{17}
-1.09	.28	-.11	-.69	-1.48	1.02	.57	-1.18	-1.04	.64	-.31	.20	-.32	-.92	.63	-.23	-.22

δ_{18}	δ_{19}	δ_{20}	δ_{21}	δ_{22}	δ_{23}	δ_{24}	δ_{25}	δ_{26}	δ_{27}	δ_{28}	δ_{29}	δ_{30}	δ_{31}	δ_{32}	δ_{33}	δ_{34}
-1.10	-.35	.22	.04	-.45	-.75	.46	-.29	.03	.57	-.01	-1.98	-.81	1.08	-.14	1.14	1.85

δ_{35}	δ_{36}	δ_{37}	δ_{38}	δ_{39}	δ_{40}	δ_{41}	δ_{42}	δ_{43}	δ_{44}	δ_{45}	δ_{46}	δ_{47}	δ_{48}	δ_{49}	δ_{50}	δ_{51}
.50	.29	.36	.29	-.20	1.59	.72	-1.00	.64	-.47	2.37	-.50	1.29	-.69	.18	-.33	-.31

δ_{52}	δ_{53}	δ_{54}	δ_{55}	δ_{56}	δ_{57}	δ_{58}	δ_{59}	δ_{60}	δ_{61}	δ_{62}	δ_{63}	δ_{64}	δ_{65}	δ_{66}
.20	.26	.33	-.28	.32	.73	.86	-1.76	1.18	-.17	-.54	.74	-.03	.83	.89

Again, all items are taken into account, and 100 individuals from each of the now 56 countries are drawn. A comparison between the estimated difficulties of the OECD dataset (Table 4.1) and all countries (Table 4.8) shows slight differences between the values, but overall similar results.

Table 4.9: Estimated coefficients of the reading items of the PISA data (2006) for all countries. The coefficients $\gamma_{33,1}$ to $\gamma_{66,1}$ are cross-loadings.

$\gamma_{1,1}$	$\gamma_{2,1}$	$\gamma_{3,1}$	$\gamma_{4,1}$	$\gamma_{5,1}$	$\gamma_{6,1}$	$\gamma_{7,1}$	$\gamma_{8,1}$	$\gamma_{9,1}$	$\gamma_{10,1}$	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{13,1}$	$\gamma_{14,1}$
.75	.77	1.05	1.13	.67	.73	.99	.91	.90	.35	.83	1.15	1.28	1.04
$\gamma_{15,1}$	$\gamma_{16,1}$	$\gamma_{17,1}$	$\gamma_{18,1}$	$\gamma_{19,1}$	$\gamma_{20,1}$	$\gamma_{21,1}$	$\gamma_{22,1}$	$\gamma_{33,1}$	$\gamma_{34,1}$	$\gamma_{35,1}$	$\gamma_{36,1}$	$\gamma_{37,1}$	$\gamma_{38,1}$
1.04	.85	.77	.91	.76	.60	.95	1.17	-.37	-.58	.12	.42	.17	.15
$\gamma_{39,1}$	$\gamma_{40,1}$	$\gamma_{41,1}$	$\gamma_{42,1}$	$\gamma_{43,1}$	$\gamma_{44,1}$	$\gamma_{45,1}$	$\gamma_{46,1}$	$\gamma_{47,1}$	$\gamma_{48,1}$	$\gamma_{49,1}$	$\gamma_{50,1}$	$\gamma_{51,1}$	$\gamma_{52,1}$
.42	-.26	-.35	.14	.26	.46	-.28	.18	-.39	.20	.04	.26	.14	-.08
$\gamma_{53,1}$	$\gamma_{54,1}$	$\gamma_{55,1}$	$\gamma_{56,1}$	$\gamma_{57,1}$	$\gamma_{58,1}$	$\gamma_{59,1}$	$\gamma_{60,1}$	$\gamma_{61,1}$	$\gamma_{62,1}$	$\gamma_{63,1}$	$\gamma_{64,1}$	$\gamma_{65,1}$	$\gamma_{66,1}$
-.17	.01	.23	.15	.07	-.12	.30	-.05	.23	.24	.07	.20	.19	-.07

The coefficients of reading and math abilities in Tables 4.9 and 4.10 and the coefficients of their interaction in Table 4.11, are very similar to the estimates of the OECD data (Tables 4.2, 4.3, and 4.4), as well. Since the estimated correlation between the latent variables for all countries is also close to the value for the OECD countries, it can be assumed that the model itself is stable and is not affected by changing the range of the inspected countries.

Table 4.10: Estimated coefficients of the math items of the PISA data (2006) for all countries. The coefficients $\gamma_{33,2}$ to $\gamma_{66,2}$ are cross-loadings.

$\gamma_{23,2}$	$\gamma_{24,2}$	$\gamma_{25,2}$	$\gamma_{26,2}$	$\gamma_{27,2}$	$\gamma_{28,2}$	$\gamma_{29,2}$	$\gamma_{30,2}$	$\gamma_{31,2}$	$\gamma_{32,2}$	$\gamma_{33,2}$	$\gamma_{34,2}$	$\gamma_{35,2}$	$\gamma_{36,2}$	$\gamma_{37,2}$
.66	.94	.85	.73	.96	.64	.57	.80	.98	.55	1.40	1.44	.70	.67	.46
$\gamma_{38,2}$	$\gamma_{39,2}$	$\gamma_{40,2}$	$\gamma_{41,2}$	$\gamma_{42,2}$	$\gamma_{43,2}$	$\gamma_{44,2}$	$\gamma_{45,2}$	$\gamma_{46,2}$	$\gamma_{47,2}$	$\gamma_{48,2}$	$\gamma_{49,2}$	$\gamma_{50,2}$	$\gamma_{51,2}$	$\gamma_{52,2}$
.57	.96	.77	.88	.31	.86	.68	.78	.70	1.38	.48	.92	.59	.65	.63
$\gamma_{53,2}$	$\gamma_{54,2}$	$\gamma_{55,2}$	$\gamma_{56,2}$	$\gamma_{57,2}$	$\gamma_{58,2}$	$\gamma_{59,2}$	$\gamma_{60,2}$	$\gamma_{61,2}$	$\gamma_{62,2}$	$\gamma_{63,2}$	$\gamma_{64,2}$	$\gamma_{65,2}$	$\gamma_{66,2}$	
.75	.89	.44	.58	.84	.81	.10	1.05	.78	.75	.98	.72	.59	.74	

The variance of the random intercept, however, is much higher, when all countries are taken into account than for the OECD countries. The absolute values only differ by 0.07, but that can be regarded as a significant increase for a variance of a random intercept, especially in comparison to the fixed variances of the latent variables to 1. As expected, this result shows that there is more variation between countries, if not only the OECD countries are considered.

Table 4.11: Estimated coefficients of the interaction effects, the correlation between the latent variables, and the variance of the random intercept of the PISA data (2006) for all countries. The coefficients $\gamma_{33,1}$ to $\gamma_{66,1}$ are cross-loadings.

$\Omega_{33}^{(2,1)}$	$\Omega_{34}^{(2,1)}$	$\Omega_{35}^{(2,1)}$	$\Omega_{36}^{(2,1)}$	$\Omega_{37}^{(2,1)}$	$\Omega_{38}^{(2,1)}$	$\Omega_{39}^{(2,1)}$	$\Omega_{40}^{(2,1)}$	$\Omega_{41}^{(2,1)}$	$\Omega_{42}^{(2,1)}$	$\Omega_{43}^{(2,1)}$	$\Omega_{44}^{(2,1)}$
.09	.20	.00	-.01	.12	.00	.02	.12	.16	-.08	-.01	-.02
$\Omega_{45}^{(2,1)}$	$\Omega_{46}^{(2,1)}$	$\Omega_{47}^{(2,1)}$	$\Omega_{48}^{(2,1)}$	$\Omega_{49}^{(2,1)}$	$\Omega_{50}^{(2,1)}$	$\Omega_{51}^{(2,1)}$	$\Omega_{52}^{(2,1)}$	$\Omega_{53}^{(2,1)}$	$\Omega_{54}^{(2,1)}$	$\Omega_{55}^{(2,1)}$	$\Omega_{56}^{(2,1)}$
.18	.02	.12	-.07	.03	-.04	.03	.09	.10	.15	-.01	.06
$\Omega_{57}^{(2,1)}$	$\Omega_{58}^{(2,1)}$	$\Omega_{59}^{(2,1)}$	$\Omega_{60}^{(2,1)}$	$\Omega_{61}^{(2,1)}$	$\Omega_{62}^{(2,1)}$	$\Omega_{63}^{(2,1)}$	$\Omega_{64}^{(2,1)}$	$\Omega_{65}^{(2,1)}$	$\Omega_{66}^{(2,1)}$	ρ	σ^2
.07	.10	-.09	.12	.01	.01	.03	-.01	-.03	.02	.72	.10

Possible interactions between the latent variables, that effect the solving probability of an item, are confirmed for all data as well. In combination with high loadings of reading ability on math items, this might indicate that solving a math item also depends on the ability to understand the item on a linguistic level.

In the reduced OECD model, the interaction effects could not be properly estimated. Therefore, no additional reduced model for all of the countries was estimated.

Overall, the analysis of PISA data with MINoLEM shows its applicability to real data and how it could help to examine complex datasets. The results suggest that the probability of answering a math item in PISA depends not only on math ability, but also on reading ability. Additionally, the influence of the math ability might be intermediated by the reading ability. More in depth calculations are needed to confirm this theory.

Chapter 5

Discussion and Conclusion

In this chapter, the main theoretical results of the thesis, as well as the results of the simulations, and the application to the PISA data will be summarized. The thesis will also be set in relation to the current state of research on latent variable models. Subsequently, possible extensions of MINoLEM and future research will be discussed.

5.1 Context and Summary

The estimation of latent variable models is an important research topic in many different fields and is therefore conducted from several different perspectives. One widely used approach are SEMs, which assume that (continuous) manifest variables are influenced by (continuous) latent variables. Their main goal is to analyze the relationships between the latent variables.

If the manifest variables are not continuous, and the researcher is mostly interested in the effect of latent variables, and other factors on the manifest variables, IRT provides a framework for investigating these relationships. As was shown in this thesis, there is a very close connection between IRT and SEM, and their distinction lies more in perspective than in their mathematical estimation approach. Many IRT models – as well as some multilevel models – can therefore also be formulated and estimated in a SEM setting.

Models based on generalized linear models, such as GLLAMM and GLLVM, incorporate many SEM and IRT models (including hierarchical structures) and therefore build a partly unifying framework.

The model in this thesis is presented as an IRT model but can potentially be described by using all of the aforementioned frameworks. Furthermore, MINoLEM can extend some aspects of these frameworks, since it includes a hierarchical structure as well as nonlinear influences of latent variables. Many existing models only allow for one of those properties.

Not only can frameworks vary but also the approach with which a model is estimated. A short overview of such methods was given, while two of the most frequently used estimation procedures – the classic EM and the Metropolis-Hating Robins-Monroe algorithm – were discussed in more detail. The decision in favor of the EM algorithm was based on simulation studies suggesting that MHRM, while very efficient, might be slightly less accurate. MINoLEM includes two complex aspects – a hierarchical structure and nonlinear latent variable effects – that each on its own require a substantial number of observations to be estimated correctly. Accordingly, the construction of an estimation procedure was focused on accuracy. The GHQ can – theoretically – be made as exact as the applier wishes by increasing the number of quadrature points. Therefore, the GHQ was favored over Bayesian sampling methods or the adaptive GHQ.

Thus, an EM algorithm was derived that allows for the estimation of a multilevel IRT model with nonlinear latent variable effects. The maximization of the complete data log-likelihood, with respect to the item and variance-covariance parameters conditional on the posterior probability, is carried out with a BFGS algorithm. This is an iterative quasi-Newton method for nonlinear problems, which uses the Hessian matrix to update the solution. To increase the efficiency and accuracy of the maximization, the analytic derivations are supplied to the algorithm.

Furthermore, the choice of starting values plays an important role in the speed of convergence and in efforts to avoid local maxima. To improve both, heuristics were developed that outperform the common choice of setting starting points to fixed values.

Finally, MINoLEM was applied to data. First, extensive simulations studies showed that MINoLEM is capable of estimating a single-level model with latent interaction effects equally well as the R package `ltm`. Extending the `ltm` model by the estimation of the correlation between the latent variables demonstrated that the additional parameter does not affect the results much. The difficulties are estimated consistently, and the coefficients of the latent variables are estimated with small variances and biases that partially show consistency. The interaction effects were be estimated with low RMSE, bias, and also variance proving that MINoLEM works as intended.

Furthermore, it was shown that the estimation of loadings, set to 0, does not pose a problem, so that specific model assumptions can be tested. Data, simulated without hierarchical structure, were correctly identified by MINoLEM, by estimating the variance of the random intercept to zero.

Subsequently, MINoLEM was applied to simulated data with two levels and interaction effects. The results indicated consistency for the estimation of the difficulties, the loadings of the latent variables, and the correlation between the latent variables. The consistency of the estimation of the random intercept depended on the way the number of individuals was increased. If the number of clusters was fixed, but the number of individuals per cluster increased, the bias slightly increased. If the number of individuals per cluster was fixed, but the number of clusters increased, the bias slightly decreased. An explanation might be that adding more people per cluster gives more weight to the estimation of the item parameters, while adding more clusters increases the information about the differences between the clusters. Further studies are needed to investigate this behavior and improve the consistency.

The comparison of MINoLEM with the `mirt` showed that both approaches produce similar results. The R package `mirt` has some expected bias for those items that depend on the interaction of the latent variables, since the interaction was not estimated in `mirt` resulting in misspecification of the model.

Overall, it was be demonstrated that MINoLEM can add to the existing literature and software by estimating a multilevel IRT model with nonlinear effects of the latent variables.

The applicability and usefulness for real datasets was shown by reanalyzing PISA data with MINoLEM. The results indicated that the items to measure the mathemetics ability of students might also depend on the student's reading ability, as well as on the interaction of both latent traits. The estimation of the variance of a random intercept for the different countries (OECD and other) implied that a multilevel model might be suitable for the data.

The implementation is not efficient enough yet, to bootstrap the standard errors for a complete PISA dataset with all items. Further calculations are needed to explore, to what extent MINoLEM could improve the estimation of PISA data. Further limitations of MINoLEM are addressed in the following section.

5.2 Future Extensions and Research

There are still aspects that can be extended in the future to improve the model's applicability. For now, only dichotomous items can be analyzed, but polytomous items could be included in the future, as it is done in the GLLAMM or GLLVM frameworks. The extension to polytomous items also opens the door to the analysis of Generalized Partial Credit or Graded response models. More possible models can be found in van der Linden (2016a, 2016b, 2016c) and Kelava, Noventa, and Robitzsch (2020).

The coefficients of the latent variables and their interactions do not vary between clusters. A random slope model can be built, in which the coefficients follow a distribution that depends on the cluster structure. It can be expected, however, that both of these extensions would require a larger number of items and observations to be estimated accurately.

Furthermore, other existing IRT models could be taken into consideration. Covariates, as well as guessing and ceiling parameters, could be added to extend the 3PL and 4PL model with a multilevel structure and nonlinear latent variable effects.

Rizopoulos and Moustaki (2008) suggested a hybrid algorithm. The EM is applied for some iterations since it converges quickly to the proximity of a solution. As soon as the improvement of the EM estimates slows down, a *common* optimization of the observed data likelihood function is applied, that converges quickly if the starting values are already close to a solution. A similar approach as in Rizopoulos and Moustaki (2008) could be investigated, where they found equality of the derivative of the *observed* data log-likelihood and the expectancy of the *complete* data log-likelihood conditional on the posterior probability of the latent variables. The latter is the main component of the EM.

In the implementation of the MINoLEM estimation, the latent variables are assumed to follow a (multivariate) normal distribution. In many cases, this is a sensible assumption, but the inclusion of other distributions might be necessary for some applications. Since the estimation is conducted using Gauss-Hermite quadrature, a first step might be the inclusion of mixtures of normal distributions, as in Bauer (2003), for example. A switch to other integral approximation methods – like MCMC sampling – might make distribution-free estimations possible. Current advances in research could also help with extensions – for example, Garcia and Ma (2016) propose an estimator for logistic models with a distribution-free random intercept.

The simulations showed that the random intercept is slightly underestimated. This could be improved, for example, by analyzing the approach by Elff, Heisig, Schaeffer, and Shikano (2020), who use restricted Maximum Likelihood estimation to achieve unbiased estimates. In simulated data with very high coefficients for the interaction terms, the estimates of those coefficients showed some bias. This can be expected for such high values, since the distinction between values becomes numerically very difficult, when the number of items and the amount of observed data is not sufficiently large. Nevertheless, an improvement in accuracy should be investigated, in order to deliver better performance for smaller datasets.

In extending the existing estimation method, different estimation paths could also be explored. As explained earlier, the EM was chosen over the MHRM, but additionally estimating the model, using this more Bayesian form of the EM, might provide additional insights. Specifically, this method might provide a more efficient estimation, while maintaining the stability.

The estimation of complex models can converge in a local minimum. This is addressed in the current approach by choosing starting values that are already close to the optimal solution. Better heuristics might be found by applying more complex methods to generate good starting points. The deterministic annealing EM algorithm by Ueda and Nakano (1998) provides – in simpler models – estimates that do not depend on the starting values. Since the starting values are calculated in an ascending manner, beginning with the simplest model, applying this approach for several iterations on less complex instances of the model might provide more accurate starting values.

Apart from examining possible extensions, MINoLEM can also be applied to different data to investigate where the new model might improve the analysis. The application of the MINoLEM estimation to a PISA dataset, for example, showed that the data indicate a potentially more complex relationship in the assessment of students' ability than is currently considered. More detailed explorations of PISA and other datasets might prove that more complicated models need to be included in such analyses - especially in the educational context. In future projects, the data from all previous PISA studies could be analyzed and compared to the official results.

With increasing computational capacities, it is to be expected that more complex data is gathered in a more complex fashion. Advanced methods are therefore needed by researchers. MINoLEM's property of simultaneously estimating a hierarchical structure and nonlinear effects of the latent variables provides the opportunity to analyze those complex datasets with additional assumptions in the structure of the data.

Appendix A

Supplementary Information

A.1 Fisher's Identity

In the context of derivatives it is worth noting that the EM algorithm can also be discussed starting with Fisher's Identity, which describes a connection between the derivative of the observed data log-likelihood and of the complete data log-likelihood

$$\frac{\partial}{\partial \boldsymbol{\omega}} \log P(y_j|\boldsymbol{\omega}) = \int \frac{\partial}{\partial \boldsymbol{\omega}} P(y_j, \xi|\boldsymbol{\omega}) P(\xi|y_j, \boldsymbol{\omega}) d\xi$$

where y_j is the observation belonging to a person j with ability ξ_j and $\boldsymbol{\omega}$ are again the parameters of the model behind $P(y_j|\xi, \boldsymbol{\omega})$. Essentially, Fisher's Identity assures equality if the derivatives of the individual likelihoods are considered in the final objective function (2.18) $B_O(\boldsymbol{\omega}, \boldsymbol{\omega}^k) = \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega}^k) \log L(\boldsymbol{Y}, \boldsymbol{\xi}|\boldsymbol{\omega})$ while equality without the derivatives can only be proven for

$$\log L(\boldsymbol{Y}|\boldsymbol{\omega}) = \sum_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{\xi}|Y, \boldsymbol{\omega}) \log \frac{L(\boldsymbol{Y}, \boldsymbol{\xi}|\boldsymbol{\omega})}{P(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{\omega})}$$

as in (2.10). The roots of the derived observed log-likelihood and of the expectancy of the derived complete data log-likelihood conditional on the posterior probability of the latent variable $\boldsymbol{\xi}$ are equal. This serves as a rationale for optimizing the expectancy of the complete data log-likelihood conditional on the posterior probability of the latent variable instead of the observed log-likelihood, which is the result of the EM in the objective function $B_O(\boldsymbol{\omega}, \boldsymbol{\omega}^k)$.

Fisher's Identity can also be extended to include a multilevel structure:

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\omega}} \log \left(P(\boldsymbol{Y}|\boldsymbol{\omega})\right) \\
=&\frac{\partial}{\partial \boldsymbol{\omega}} \log \left[\prod_{k=1}^{K} \int \left[\prod_{j=1}^{J_k} \int \left[\prod_{i=1}^{I} P(Y_{ijk}, \boldsymbol{\xi}, u|\boldsymbol{\omega}) \right] d\boldsymbol{\xi} \right] du \right] \qquad \text{(A.1)} \\
=&\sum_{k=1}^{K} \frac{\partial}{\partial \boldsymbol{\omega}} \log \left[\int \left[\prod_{j=1}^{J_k} \int \left[\prod_{i=1}^{I} P(Y_{ijk}, \boldsymbol{\xi}, u|\boldsymbol{\omega}) \right] d\boldsymbol{\xi} \right] du \right] \qquad \text{(A.2)} \\
=&\sum_{k=1}^{K} \frac{\frac{\partial}{\partial \boldsymbol{\omega}} \left[\int \left[\prod_{j=1}^{J_k} \int \left[\prod_{i=1}^{I} P(Y_{ijk}, \boldsymbol{\xi}, u|\boldsymbol{\omega}) \right] d\boldsymbol{\xi} \right] du \right]}{\int \left[\prod_{j=1}^{J_k} \int \left[\prod_{i=1}^{I} P(Y_{ijk}, \boldsymbol{\xi}, u|\boldsymbol{\omega}) \right] d\boldsymbol{\xi} \right] du} \qquad \text{(A.3)} \\
\overset{Leibniz}{=}&\sum_{k=1}^{K} \frac{\int \frac{\partial}{\partial \boldsymbol{\omega}} P(\boldsymbol{Y}_k, u|\boldsymbol{\omega}) du}{P(Y_k|\boldsymbol{\omega})} \qquad \text{(A.4)} \\
=&\sum_{k=1}^{K} \int \frac{\partial}{\partial \boldsymbol{\omega}} \left[\log P(\boldsymbol{Y}_k, u|\boldsymbol{\omega})\right] \frac{P(\boldsymbol{Y}_k, u|\boldsymbol{\omega})}{P(\boldsymbol{Y}_k|\boldsymbol{\omega})} du \qquad \text{(A.5)} \\
=&\sum_{k=1}^{K} \int \frac{\partial}{\partial \boldsymbol{\omega}} \left[\log P(\boldsymbol{Y}_k, u|\boldsymbol{\omega})\right] P(u|\boldsymbol{Y}_k, \boldsymbol{\omega}) du \qquad \text{(A.6)} \\
\overset{\text{(A.1)}-\text{(A.6)}}{=}&\sum_{k=1}^{K} \int \left[\sum_{j=1}^{J_k} \int \left[\frac{\partial}{\partial \boldsymbol{\omega}} \left[\log \left(P(\boldsymbol{Y}_{jk}, \boldsymbol{\xi}, u|\boldsymbol{\omega})\right) \right] P(\boldsymbol{\xi}|\boldsymbol{Y}_{jk}, \boldsymbol{\omega}) \right] d\boldsymbol{\xi} P(u|\boldsymbol{Y}_k, \boldsymbol{\omega}) \right] du
\end{aligned}
$$

In (A.5), the property $f'(x) = (log(f(x)))' f(x)$ of the logarithm is used. In the last line, all the steps from Equations (A.1) to (A.6) are done in the same manner, but for $\frac{\partial}{\partial \boldsymbol{\omega}} \left[\log P(\boldsymbol{Y}_k, u|\boldsymbol{\omega})\right]$ instead of $\frac{\partial}{\partial \boldsymbol{\omega}} \log \left(P(\boldsymbol{Y}|\boldsymbol{\omega})\right)$.

A.2 Graph of Convergence

In this section, the development of an estimation of a MINoLEM is depicted in a plot. It is an exemplary course of estimation and can be observed similarly in most of the simulated datasets in Chapter 3, as the results show.

One dataset was simulated as in the multilevel simulation study with $N_C = N_S = 100$. The difficulties, named 'd_01' to 'd_10' in the graph, for the 10 items were, once again, set to

$$\boldsymbol{\delta} = \begin{pmatrix} 1 & -1.2 & -0.2 & 0.6 & 1.2 & -0.6 & 0.2 & -1 & 0 & -0.4 \end{pmatrix}^t ,$$

and a two-dimensional latent variable with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_2})$ was assumed. Each latent variable had four items that depended only on them. The last two items were affected by both latent variable dimensions. In the plot, the coefficients are named 'g_01,1', for example, which stands for $\gamma_{1,1}$, and so forth. The coefficients are given by

$$\boldsymbol{\gamma} = \begin{pmatrix} 1 & 0.5 & 0.55 & 1.2 & 0 & 0 & 0 & 0 & 0.45 & 1.1 \\ 0 & 0 & 0 & 0 & 1 & 1.15 & 0.95 & 0.6 & 1.05 & 0.65 \end{pmatrix}^t .$$

Items nine and ten are also influenced by the interaction of the latent variables

$$\boldsymbol{\Omega}_9 = \begin{pmatrix} 0 & 0 \\ 0.1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega}_{10} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} ,$$

which are named 'o_01' and 'o_02', respectively, in the plot. The variance of the random intercept, named 's' in the plot, was set to $\sigma^2 = 0.125$, and the correlation between the latent variables, named 'p', to $\rho = 0.3$. In the figures, those true values are depicted as dotted lines, while the continuous lines represent the estimates. The convergence criterion was met after 16 iterations. The estimated values for each iteration are given in Tables A.1 and A.2.

The upper plot A in Figure A.1 shows the values of the difficulty estimates in each iteration. It can be observed that the difficulties are quickly estimated very close to the true values – in this simulation even after the first iteration. The coefficients of the latent variables in the lower plot B in A.1 are continuously improved with each iteration. After approximately ten iterations all coefficient estimates are very close to the true values and don't improve significantly. The estimates of the coefficients converge nicely to the true values.

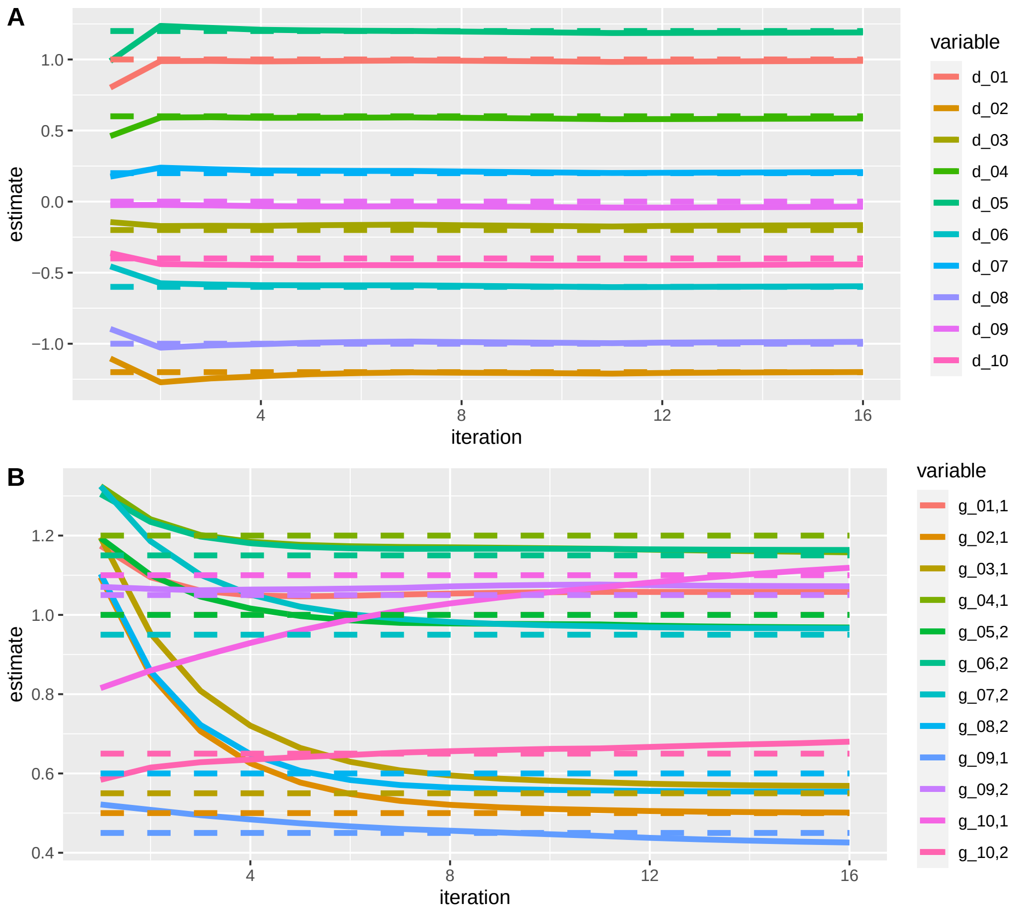

Figure A.1: Plot of convergence of the difficulties (A) and the latent variable coefficients (B) in exemplary dataset. The dotted lines represent the true values. The continuous lines represent the estimates in each iteration. The difficulties δ_1 to δ_{10} are named ‘d_01’ to ‘d_10’. The loadings of the latent variables are named ‘g_01,1’, for example, which stands for $\gamma_{1,1}$, and so forth.

The estimation of the correlation, named ‘p’, between the latent variables in plot A in Figure A.2 improves in each iteration and converges to the true value. In this simulation, the variance of the random intercept, named ‘s’, in plot B in Figure A.2 is already estimated well by the heuristic for the starting value. The estimate in the last iteration is slightly improved.

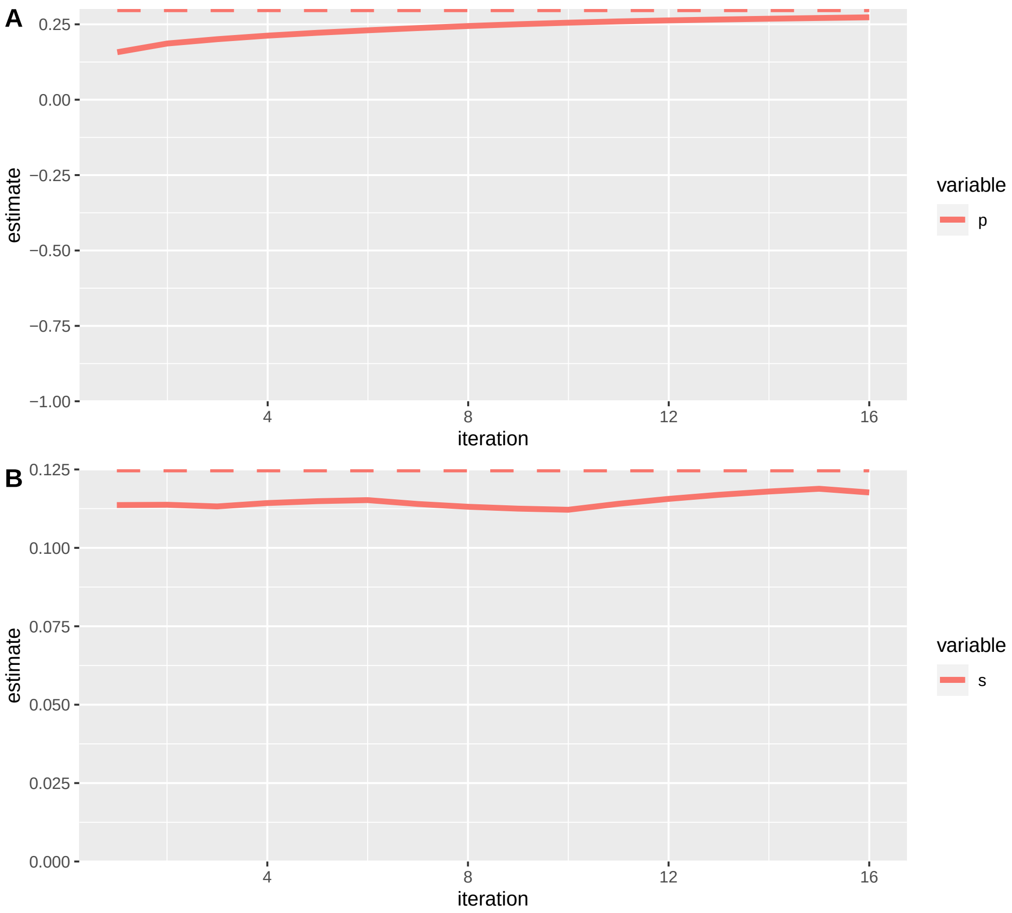

Figure A.2: Plot of convergence of the correlation between the latent variables (A) and the variance of the random intercept (B) in exemplary dataset. The dotted lines represent the true values. The continuous lines represent the estimates in each iteration. The correlation ρ is named ‘p’ in the plot. The variance of the random intercept σ^2 is named ‘s’.

The estimates of the interaction effects in plot A in Figure A.3 improve in each iteration and converge to the true values, as the coefficient estimates did. In the last iteration the estimates are very close to the true values.

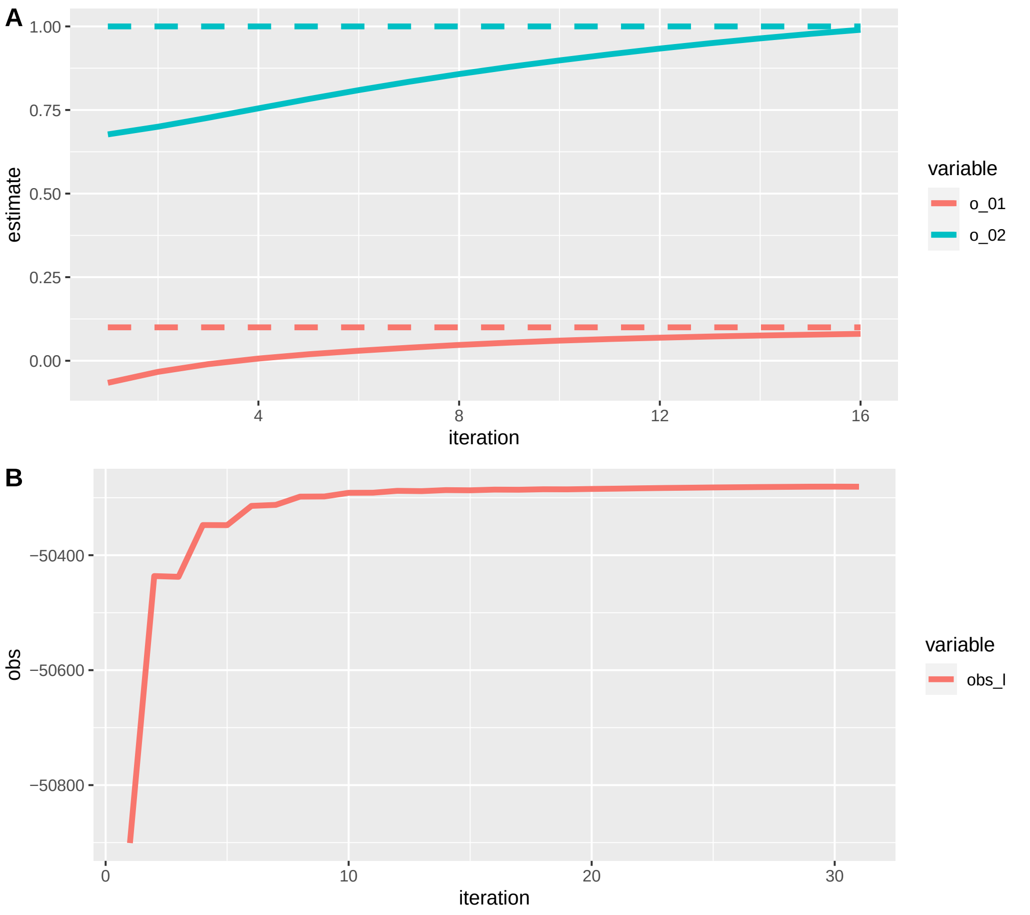

Figure A.3: Plot of convergence of the interaction coefficients (A) and of the values of the observed data log-likelihood (B) in exemplary dataset. The graph B shows values for $2 \cdot 16 = 32$ iterations, since all the log-likelihood values are presented after the estimation of the item parameters and after the estimation of the variance of the random intercept. The dotted lines represent the true values. The continuous lines represent the estimates in each iteration. The interaction coefficients $\boldsymbol{\Omega}_9$ and $\boldsymbol{\Omega}_{10}$ are named 'o_01' and 'o_02', respectively. The observed data log-likelihood is named 'obs_l'.

As expected, the values of the observed data log-likelihood $\log P(\boldsymbol{Y}|\boldsymbol{\omega})$, named 'obs_l', in plot B in Figure A.3 increase in each iteration. The graph B shows values for $2 \cdot 16 = 32$ iterations, since all the log-likelihood values are presented that result after each estimation of the item parameters and after each estimation of the variance of the random intercept. It can be observed that the log-likelihood does not seem to rise much after iteration five. However, the improvements of the estimates are still significant after that. That shows that the objective function is sensible to improvements of the estimates, even if the changes in the objective function are minimal.

Overall, it can be noted that good estimates are obtained after a few iterations. Especially, the item difficulties are quickly estimated accurately. The estimates of the interaction effects are the only ones that significantly improve until the last iteration. This indicates that the convergence criterion works well.

Table A.1: Results in each iteration of exemplary multilevel dataset with nonlinear latent variable effects – part 1.

It	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}	$\gamma_{1,1}$	$\gamma_{2,1}$	$\gamma_{3,1}$
1	.802	-1.104	-.144	.461	.989	-.456	.176	-.896	-.025	-.362	1.174	1.095	1.195
2	.989	-1.271	-.172	.591	1.236	-.575	.239	-1.028	-.024	-.440	1.095	.848	.953
3	.990	-1.244	-.170	.594	1.222	-.583	.228	-1.012	-.027	-.444	1.061	.707	.809
4	.986	-1.229	-.171	.589	1.209	-.588	.219	-1.003	-.032	-.447	1.049	.626	.721
5	.988	-1.214	-.166	.590	1.205	-.589	.217	-.993	-.034	-.448	1.047	.577	.664
6	.991	-1.206	-.164	.591	1.202	-.589	.216	-.988	-.034	-.447	1.048	.548	.629
7	.993	-1.202	-.162	.592	1.200	-.589	.215	-.985	-.033	-.447	1.051	.531	.608
8	.991	-1.204	-.166	.590	1.196	-.592	.211	-.988	-.034	-.447	1.054	.521	.595
9	.989	-1.206	-.169	.586	1.193	-.595	.208	-.991	-.037	-.449	1.056	.515	.587
10	.986	-1.208	-.172	.583	1.189	-.599	.204	-.994	-.039	-.450	1.057	.510	.582
11	.983	-1.210	-.175	.580	1.186	-.602	.201	-.996	-.042	-.449	1.058	.508	.578
12	.984	-1.206	-.171	.580	1.186	-.601	.203	-.992	-.043	-.449	1.058	.505	.574
13	.986	-1.204	-.170	.581	1.187	-.600	.204	-.991	-.041	-.447	1.058	.504	.572
14	.987	-1.202	-.168	.583	1.188	-.598	.205	-.989	-.039	-.445	1.058	.503	.570
15	.989	-1.201	-.167	.584	1.189	-.597	.206	-.988	-.037	-.443	1.058	.502	.569
16	.990	-1.199	-.166	.585	1.190	-.596	.207	-.987	-.036	-.442	1.058	.501	.569

Table A.2: Results in each iteration of exemplary multilevel dataset with nonlinear latent variable effects – part 2.

It	$\gamma_{4,1}$	$\gamma_{9,1}$	$\gamma_{10,1}$	$\gamma_{5,2}$	$\gamma_{6,2}$	$\gamma_{7,2}$	$\gamma_{8,2}$	$\gamma_{9,2}$	$\gamma_{10,2}$	$\Omega_9^{(2,1)}$	$\Omega_{10}^{(2,1)}$	ρ	σ^2
1	1.325	.522	.815	1.194	1.304	1.325	1.103	1.071	.584	-.066	.677	.158	.114
2	1.241	.508	.860	1.101	1.235	1.186	.858	1.066	.615	-.033	.700	.186	.114
3	1.201	.495	.895	1.046	1.198	1.101	.723	1.062	.628	-.010	.727	.201	.113
4	1.185	.484	.929	1.016	1.181	1.051	.649	1.064	.635	.007	.755	.213	.114
5	1.177	.474	.961	.997	1.172	1.021	.607	1.065	.641	.019	.783	.222	.115
6	1.173	.467	.988	.986	1.168	1.002	.583	1.066	.647	.030	.809	.230	.115
7	1.171	.460	1.011	.980	1.167	.990	.571	1.068	.653	.039	.834	.238	.114
8	1.170	.456	1.029	.978	1.167	.982	.565	1.072	.656	.047	.858	.244	.113
9	1.169	.451	1.044	.977	1.167	.977	.561	1.074	.659	.054	.879	.250	.113
10	1.168	.447	1.057	.977	1.167	.974	.559	1.076	.662	.060	.898	.255	.112
11	1.167	.442	1.069	.976	1.166	.971	.557	1.076	.663	.065	.917	.260	.114
12	1.164	.438	1.081	.973	1.165	.969	.556	1.075	.667	.069	.934	.263	.116
13	1.163	.434	1.093	.971	1.165	.967	.555	1.074	.670	.072	.950	.266	.117
14	1.161	.431	1.102	.970	1.164	.967	.554	1.073	.673	.075	.964	.269	.118
15	1.159	.428	1.111	.969	1.164	.966	.554	1.073	.676	.078	.978	.271	.119
16	1.158	.426	1.119	.968	1.164	.966	.554	1.072	.680	.080	.990	.273	.118

Appendix B

Additional Tables

B.1 Tables for Single-Level Data with Estimated Correlation

Table B.1: RMSE, bias, and variance of the difficulties for simulation of single-level model and estimated correlation between the latent variables of $\rho = 0.3$. The interaction terms have medium values. E_M = Estimation with MINoLEM.

N		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.157	.024	.024	.126	-.019	.015	.102	-.011	.010	.139	.017	.019	.159	.023	.025
1000	E_M	.097	.014	.009	.085	-.007	.007	.067	.003	.005	.092	.014	.008	.102	.003	.010
2000	E_M	.069	.010	.005	.062	-.004	.004	.049	-.000	.002	.066	.007	.004	.070	.007	.005
5000	E_M	.042	.007	.002	.033	.003	.001	.029	-.002	.001	.041	-.001	.002	.042	.001	.002

N_C		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.118	.002	.014	.111	.021	.012	.120	-.006	.014	.140	.031	.019	.177	-.015	.031
1000	E_M	.083	-.002	.007	.076	.002	.006	.080	-.002	.006	.091	.000	.008	.119	-.019	.014
2000	E_M	.060	-.005	.004	.052	-.001	.003	.065	-.004	.004	.060	.002	.004	.083	-.013	.007
5000	E_M	.043	.000	.002	.032	-.001	.001	.034	-.000	.001	.038	-.003	.001	.047	-.013	.002

Table B.2: RMSE, bias, and variance of the coefficients of the latent variables for simulation of single-level model and estimated correlation between the latent variables of $\rho = 0.3$. The interaction terms have medium values. E_M = Estimation with MINoLEM.

N		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.266	.056	.068	.194	.029	.037	.169	.022	.028	.322	.052	.101
1000	E_M	.185	.027	.034	.129	.013	.016	.106	-.004	.011	.198	.018	.039
2000	E_M	.129	.023	.016	.103	.003	.011	.079	.007	.006	.143	.018	.020
5000	E_M	.073	.016	.005	.052	.006	.003	.055	.007	.003	.086	.009	.007

N		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.264	.069	.065	.444	.062	.193	.254	.018	.064	.275	.047	.010
1000	E_M	.172	.038	.028	.248	-.003	.062	.174	-.000	.030	.181	.015	.010
2000	E_M	.108	.006	.012	.177	-.036	.030	.114	.009	.013	.113	.017	.010
5000	E_M	.070	.009	.005	.116	-.028	.013	.072	.001	.005	.078	.008	.010

N		$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.229	.012	.052	.183	-.009	.033	.300	.091	.082	.337	.042	.112
1000	E_M	.140	.008	.020	.114	.007	.013	.214	.070	.041	.196	-.010	.039
2000	E_M	.110	.022	.012	.088	.007	.008	.145	.039	.019	.111	-.006	.012
5000	E_M	.063	.000	.004	.049	.002	.002	.084	.022	.007	.087	-.009	.008

Table B.3: RMSE, bias, and variance of the difficulties a for simulation of single-level model and estimated correlation between the latent variables of $\rho = 0.3$. The interaction terms have high values. E_M = Estimation with MINoLEM.

N		δ_1			δ_2			δ_3			δ_4			δ_5		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.162	.029	.026	.124	-.022	.015	.103	-.011	.010	.142	.025	.019	.162	.028	.025
1000	E_M	.097	.017	.009	.085	-.008	.007	.067	.002	.005	.095	.018	.009	.103	.010	.011
2000	E_M	.069	.011	.005	.061	-.005	.004	.049	-.000	.002	.067	.013	.004	.071	.011	.005
5000	E_M	.043	.009	.002	.033	.002	.001	.029	-.002	.001	.041	.004	.002	.044	.005	.002

N		δ_6			δ_7			δ_8			δ_9			δ_{10}		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.120	-.005	.014	.112	.021	.012	.118	-.008	.014	.200	.056	.037	.178	.006	.032
1000	E_M	.083	-.006	.007	.077	.003	.006	.079	-.001	.006	.138	.051	.017	.132	-.017	.017
2000	E_M	.061	-.009	.004	.052	.000	.003	.065	-.004	.004	.092	.032	.007	.086	-.007	.007
5000	E_M	.044	-.002	.002	.032	.000	.001	.035	-.000	.001	.070	.028	.004	.054	-.008	.003

Table B.4: RMSE, bias, and variance of the coefficients of the latent variables for simulation of single-level model and estimated correlation between the latent variables of $\rho = 0.3$. The interaction terms have high values. E_M = Estimation with MINoLEM.

N		$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.263	.067	.064	.196	.040	.037	.163	.025	.026	.327	.088	.099
1000	E_M	.173	.036	.029	.125	.021	.015	.107	-.001	.011	.198	.041	.037
2000	E_M	.123	.023	.015	.100	.009	.010	.078	.010	.006	.150	.047	.020
5000	E_M	.078	.023	.006	.052	.010	.003	.053	.009	.003	.099	.031	.009

N		$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.332	-.007	.110	.428	-.162	.157	.261	.028	.067	.313	.084	.011
1000	E_M	.231	-.038	.052	.324	-.138	.086	.177	.018	.031	.197	.039	.011
2000	E_M	.175	-.061	.027	.285	-.158	.057	.116	.016	.013	.134	.041	.011
5000	E_M	.129	-.056	.014	.247	-.147	.039	.078	.010	.006	.089	.025	.011

N		$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
		rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
500	E_M	.245	.024	.060	.178	-.003	.032	.361	-.102	.120	.344	-.073	.113
1000	E_M	.142	.015	.020	.116	.004	.013	.245	-.041	.058	.276	-.120	.062
2000	E_M	.118	.028	.013	.087	.007	.008	.220	-.095	.039	.185	-.085	.027
5000	E_M	.067	.010	.004	.051	.003	.003	.178	-.083	.025	.180	-.101	.022

B.2 Tables for Simulation of Starting Values

Table B.5: RMSE, bias, and variance of the difficulties of simulation of starting values. Data was simulated for multilevel model ($\sigma^2 = 0.125$) and estimated correlation between the latent variables of $\rho = 0.3$. The number of individuals per cluster is fixed to $N_S = 100$. N_C = Number of clusters. The column 'fix' indicates the difference between the commonly chosen fixed starting value and the true value in the simulation.

N_C	δ_1			δ_2			δ_3			δ_4			δ_5		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.262	-.181	.036	.140	.091	.012	.060	.019	.003	.202	-.138	.022	.306	-.214	.048
100	.253	-.177	.033	.132	.089	.009	.048	.021	.002	.196	-.136	.020	.306	-.215	.047
200	.266	-.187	.036	.125	.086	.008	.037	.017	.001	.204	-.143	.021	.308	-.217	.048
fix		1			-1.2			-.2			.6			1.2	

N_C	δ_6			δ_7			δ_8			δ_9			δ_{10}		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.198	.136	.021	.072	-.035	.004	.145	.094	.012	.045	-.005	.002	.078	.044	.004
100	.195	.136	.020	.058	-.032	.002	.146	.099	.012	.033	-.003	.001	.068	.043	.003
200	.187	.131	.018	.057	-.036	.002	.131	.091	.009	.026	-.007	.001	.058	.037	.002
fix		-.6			.2			-1			0			-.4	

Table B.6: RMSE, bias, and variance of the loadings of the latent variables of simulation of starting values. Data was simulated for multilevel model ($\sigma^2 = 0.125$) and estimated correlation between the latent variables of $\rho = 0.3$. The number of individuals per cluster is fixed to $N_S = 100$. N_C = Number of clusters. The column 'fix' indicates the difference between the commonly chosen fixed starting value and the true value in the simulation.

N	$\gamma_{1,1}$			$\gamma_{2,1}$			$\gamma_{3,1}$			$\gamma_{4,1}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.277	.195	.039	.840	.594	.353	.870	.614	.379	.134	.092	.010
100	.278	.196	.039	.838	.593	.352	.864	.611	.374	.134	.093	.009
200	.283	.200	.040	.839	.593	.352	.866	.612	.375	.133	.093	.009
fix		0			.5			.45			-.2	

N_C	$\gamma_{9,1}$			$\gamma_{10,1}$			$\gamma_{5,2}$			$\gamma_{6,2}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.210	.143	.024	.398	-.278	.081	.300	.212	.045	.223	.156	.025
100	.204	.142	.022	.391	-.275	.077	.298	.210	.045	.222	.156	.025
200	.206	.144	.022	.391	-.275	.077	.298	.210	.045	.222	.156	.025
fix		.55			-.1			0			-.15	

N_C	$\gamma_{7,2}$			$\gamma_{8,2}$			$\gamma_{9,2}$			$\gamma_{10,2}$		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.500	.353	.125	.690	.488	.239	.113	.063	.009	.113	-.067	.008
100	.496	.350	.123	.687	.485	.236	.098	.061	.006	.102	-.065	.006
200	.499	.353	.125	.683	.483	.234	.093	.060	.005	.095	-.064	.005
fix		.05			.4			-.05			.35	

Table B.7: RMSE, bias, and variance of the interaction coefficients of simulation of starting values. Data was simulated for multilevel model ($\sigma^2 = 0.125$) and estimated correlation between the latent variables of $\rho = 0.3$. The number of individuals per cluster is fixed to $N_S = 100$. N_C = Number of clusters. The column 'fix' indicates the difference between the commonly chosen fixed starting value and the true value in the simulation.

N_C	$\Omega_9^{(2,1)}$			$\Omega_{10}^{(2,1)}$			ρ			σ^2		
	rmse	bias	var	rmse	bias	var	rmse	bias	var	rmse	bias	var
50	.143	-.007	.021	.495	-.338	.130	.203	-.143	.021	.036	-.017	.001
100	.105	.002	.011	.468	-.325	.114	.202	-.143	.021	.031	-.018	.001
200	.070	.004	.005	.472	-.331	.113	.202	-.143	.020	.027	-.017	.000
fix		.1			1			.3			.125	

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69-81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Arminger, G., & Muthen, B. (1998). A bayesian approach to nonlinear latent variable models using the gibbs sampler and the metropolis-hastings algorithm. *Psychometrika*, *63*(3), 327-300.

Arminger, G., & Stein, P. (1997). Finite mixtures of covariance structure models with regressors. *Sociological Methods and Research*, *26*(2), 148-182.

Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures and conditional mean- and covariance-structure models. *Psychometrika*, *64*(4), 475-494.

Bartholomew, D. (1987). *Latent variable models and factor analysis: A unified approach* (1st ed.). Oxford University Press, Inc. New York.

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). John Wiley & Sons, Ltd.

Bartlett, M. S. (1935). Estimation of general ability. *Nature*, *135*(71).

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item response model. *Research Bulletin*.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135-167.

Bauer, D. J. (2005). A semiparamtric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(4), 513-535.

Bentler, P. M. (2006). EQS 6 structural equation program manual [Computer software manual]. Encino, CA: Multivariate Software, Inc..

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, *41*, 5611-575.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (chap. 17-20). Reading, Mass.: Addison-Wesley.

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4).

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Bollen, K. A. (1995). Structural equation models that are nonlinear in latent variables: A least squares estimator. *Sociological Methodology*, *25*, 223-251.

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109-121.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605-634.

Brown, M. L. (1974). Identification of the sources of significance in two-way tables. *Applied Statistics*, *23*, 405-413.

Cai, L. (2008). *A metropolis–hastings robbins–monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model* (Unpublished doctoral dissertation). Department of Psychology, University of North Carolina at Chapel Hill.

Cai, L. (2010). High dimensional exploratory item factor analysis by a metropolis-hastings robbins-monro algorithm. *Psychometrika*, *75*(1), 33-57.

Cai, L. (2017). *flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring.* Computer Software. (Chapel Hill, NC: Vector Psychomeric Group)

Cai, L., & Thissen, D. (2015). Handbook of item response theory modeling. In S. Reise & D. Revicki (Eds.), (p. 41-59). Routledge, New York and London.

Cai, L., Thissen, D., & du Toit, S. (2015). *IRTPRO.* Computer Software. (Lincolnwood, IL: Scientific Software International)

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics*, *19*(3), 293-295.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program. *ACT Research Report Series*.

Chalmers, R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. doi: 10.18637/jss.v048.i06

Chen, T., & Fienberg, S. (1976). The analysis of contingency tables with incompletely classified data. *Biometrics*, *32*, 133-144.

Chen, Y., Li, X., & Zhang, S. (2017). *Joint maximum likelihood estimation for high-dimensional exploratory item response analysis.*

Choi, Y.-J., & Asilkalkan, A. (2019). R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 168-175.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5-32.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529-569.

Dellaert, F. (2002). The expectation maximization algorithm. *Georgia Institute of Technology Report*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1-38.

Du Toit, M. (2003). *IRT from SSI.* (Lincolnwood, IL: Scientific Software International)

Elff, M., Heisig, J., Schaeffer, M., & Shikano, S. (2020). Multilevel analysis with few clusters: Improving likelihood-based methods to provide unbiased estimates and accurate inference. *British Journal of Political Science*.

Fox, J.-P., & Glas, A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, *66*(2), 271-288.

Galton, F. (1888). Co-relations and their measurement. In *Proceedings of the royal society* (Vol. 45, p. 135-145).

Garcia, T., & Ma, Y. (2016). Optimal estimator for logistic model with distribution-free random intercept. *Scandinavian Journal of Statistics*, *43*, 156-171.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Transactions on Pattern Analysis and Machine Intelligence*, *6*(6).

Goldstein, H., & McDonald, R. P. (1988). A general model fo the analysis of multilevel data. *Psychometrika*, *53*(4), 445-467.

Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*(5), 815-841.

Han, K., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement*, *38*(6), 486-498.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*, 174-194.

Harwell, M., Baker, F., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likeliood and an EM algorithm: A didactic. *Journal of Educational Statistics*, *13*(3), 243-271.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus.* Routledge.

Honaker, J., King, G., & Blackwell, M. (2011). Ameilia ii: A program for missing data. *Journal of Statistial Software*, *45*(7).

Hox, J. J., Moerbeek, M., & van de Shoot, R. (2018). *Mutlilevel analysis - techniques and applications* (Third ed.). Taylor & Francis.

IQB. (2021). *Über das IQB.* https://www.iqb.hu-berlin.de/institut/about/ (29.01.2021).

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). STEMM: A general finite mixture structural equation model. *Journal of Classification*, *14*(1), 23-50.

Jöreskog, K. G., & Yang, F. (1996). Advanced structural equation modeling. issues and techniques. In G. Macoulides & R. Schumacker (Eds.), (p. 54-88). Hillsdale: Lawrence Erlbaum.

Kamata, A. (2001). Item analysis by hierarchical linear model. *Journal of Educational Measurement*, *38*, 79-93.

Kamata, A., & Bauer, D. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136-153.

Kamata, A., & Vaughn, B. K. (2010). Handbook of advanced multilevel analysis. In (chap. Multilevel IRT Modeling). Routledge.

Karlis, D., & Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, *41*, 577-590.

Kelava, A., & Brandt, H. (2009). Estimation of nonlinear latent structural equation models using the extended unconstrained approach. *Review of Psychology*, *16*(2), 123-131.

Kelava, A., & Brandt, H. (2014). A general non-linear multilevel structural equation mixture model. *Frontiers in Psychology*, *5*, 748.

Kelava, A., Kohler, M., Krzyzak, A., & Schaffland, T. F. (2017). Nonarametric estimation of a latent variable model. *Journl of Multivariate Analysis*, *154*, 112-134.

Kelava, A., Nagengast, B., & Brandt, H. (2014). A nonlinear structural equation mixture modeling approach for non-normally distributed latent predictor variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 468-491.

Kelava, A., Noventa, S., & Robitzsch, A. (2020). Testtheorie und fragebogenkonstruktion. In H. Moosbrugger & A. Kelava (Eds.), (p. 425-447). Springer.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201-210.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*(4), 557-574.

Klein, A., & Muthen, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*(4), 647-673.

Kreiner, S., & Christensen, K. (2014). Analysis of model fit and robustness. a new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210-231.

Kuo, T.-C., & Sheng, Y. (2016). A comparison of estimation methods for a multi-unidimensional graded response IRT model. *frontiers in Psychology*.

Lambert, B. (2018). *A student's guide to bayesian statistics* (J. Seaman, Ed.). SAGE.

Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*(4), 763-772.

Lee, S.-Y. (2007). *Structural equation modeling: A bayesian approach*. John Wiley & Sons.

Li, C.-H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936-949.

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*(1), 101-122.

Liu, Q., & Pierce, D. (1994). A note on gauss-hermite quadrature. *Biometrika*, *81*.

Lord, F. M. (1952). *A theory of test scores.* Richmond, VA: Psychometric Corporation. (Retrived from: http://www.psychometrika.org/journal/online/MN07.pdf)

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Reading, MA: Addison-Wesley*.

M., Z., E., M., R., M., & R.D., B. (2003). *BILOG-MG 3.* CD-ROM. (Lincolnwood, IL: ScientificSoftware International)

Maechler, M. (2020). Rmpfr: R MPFR - multiple precision floating-point reliable [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=Rmpfr` (R package version 0.8-1)

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: evaluation of alternative estimation strategies and indicator construction. *Psychol Methods*, *9*(3), 275-300.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *Britisch Journal of Mathamatical and Statisitical Psychology*, *42*(2), 215-232.

McKinley, R. L., & Reckase, M. D. (1983). An extension of the two-parameter logistic model to the multidimenional latent space. In *Research report onr 83-2.* Iowa City, IA: The American College Testing Program.

McLachlan, G. (1988). On the choice of starting values for the EM algorithm in fitting mixture models. *The Statistician*, *37*, 417-425.

McNeish, D., & Stapleton, L. (2016). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equation modeling. *Psychological Methods*, *10*(3), 259-284.

Mislevy, R. J., & Bock, R. D. (1989). Multilevel analysis of educational data. In R. D. Bock (Ed.), (p. 57-74). San Diego, CA: Academic Press.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Mouton.

Mooijaart, A., & Bentler, P. M. (2010). An alternative approach for nonlinear latent variable models. *Structural Equaion Modeling: A Multidisciplinary Journal*, *17*(3), 357-373.

Muraki, E. (1992). A generalized partial credit model: Application of an

EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. (1992). A new inference technique for factor analyzing binary items using tetrachoric correlations. In *Annual meeting of the american educational research association, san francisco.*

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376-398.

Muthén, B. (1997). Sociological methodology. In A. Raftery (Ed.), (p. 453-481). Boston: Blackwell.

Muthén, L., & Muthén, B. (1998-2017). Mplus user's guide (Eighth ed.) [Computer software manual]. Los Angeles, CA: Muthén and Muthén.

Nader, I., Tran, U., & Voracek, M. (2015). Effects of initial values and convergence criterion in the two-parameter logistic model when estimating the latent distribution in BILOG-MG 3. *PLoS ONE*, *10*(10).

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. (2016). Bayesian prior choice in IRT estimation using MCMC and varitional bayes. *frontiers in Psychology*.

Neal, R., & Hinton, G. (1998). Learning in graphical models. In M. Jordan (Ed.), (Vol. 89, p. 355-368). Springer, Dordrecht.

OECD. (2021). https://www.oecd.org/PISA/ (28.01.2021).

of the OECD, S.-G. (2017). *PISA 2015 technical report* (Tech. Rep.). OECD.

Ping, R. (1995). A parsimonious estimating technique for interaction and quadratic latent variables. *Journal of Marketing Research*, *32*, 336-347.

Plummer, M. (2007). JAGS: A program for analysis of bayesian graphical models usiing gibbs sampling. In *Proceedings of the 3rd international* *workshop on distributed statistical computing.*

Popp, M. (2010). Viel lärm um PISA: Eine qualitative-vergleichende presseanalyse zu den reaktionen auf die PISA-studie in deutschland, Österreich, spanien und mexiko. *TranState Working Papers, No. 134, Universität Brermen, Collaborative Research Center 597 - Transformation of the State, Bremen.*

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167-190.

Rasch, G. (1960). Probabilistic modes for some intelligence and attainment tests. *Copenhagen, Denmark: Danish Institute for Educational Research.*

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (p. 321-333). Berkeley: University of California Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Sage Publications.

Reckase, M. D. (2009). *Multidimensional item response theory.* Springer.

Reise, S. P., & Reviecki, D. A. (2015). *Handbook of item response theory modeling.* Routledge.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analysis. *Journal of Statistical Software*, *17*(5), 1-25.

Rizopoulos, D., & Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, *61*, 415-438.

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test analysis modules [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=TAM` (R package version 3.5-19)

Rockwood, N. (2019). *Estimating multilevel structural equation models with random slopes for latent covariates* (Unpublished doctoral dissertation). The Ohia State University.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling.

Journal of Statistical Software, *48*(2), 1-36.

Samejima, F. (1969). *Estimation of ability using a response pattern of graded responses.* Richmond, VA: Psychometric Corporation. (Retrived from: https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf)

Schilling, S., & Bock, R. D. (2005). High dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*(3), 533-555.

Schleicher, A. (2018). *PISA 2018 - insights and interpretations.*

Shireman, E., Steinley, D., & Brusco, M. (2015). Examining the effect of initialization strategies on the performance of gaussian mixture modeling. *Bahavior Research Methods*, *49*, 282-293.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*(1), 3-31.

Sijtsma, K., & Meijer, R. R. (2007). Handbook of statistics 26: Psychometrics. In C. Rao & S. Sinharay (Eds.), (p. 719-746). Elsevier.

Sims, T. (2017). *Comparison of IRTPRO 3 and Mplus 7 for multidimensional item response item parameter and examinee ability estimation* (Dissertation). Georgia State University, https://scholarworks.gsu.edu/eps_diss/186.

Singer, J. D. (1998). Using SAS Proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and BehavioralStatistics*, *23*, 323-355.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling - multilevel, longitudinal, and structural equation models.* Chapman & Hall/CRC.

Snijder, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis - an introduction to basic and adcanced multilevel modeling.* SAGE Publications.

Song, X. Y., & Lee, S.-Y. (2007). Baysian analysis of structural equation models with multinomial variables and an application to type 2 diabetic nephropathy. *Statistics in Medicine*.

Spearman, C. (1904). General intelligence, objectivly determined and measured. *American Journal of Psychology*, *15*, 201-293.

Spirtes, P. (2001). Latent structure and causal variables. In N. Smelser & P. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 12, p. 8395-8400). Elsevier.

Stan Development Team. (2018). Stan modeling language users guide and reference manual (2.18.0 ed.) [Computer software manual]. http://mc-stan.org.

Stanley, L. (2017). *Flexible multidimensional item response theory models incorporating response styles* (Dissertation). The Ohio State University, https://etd.ohiolink.edu/!etd.send_file?accession=osu1494316298549437&disposition=inline.

StataCorp. (2019). *Stata statistical software: Release 16.* College Station, TX: StataCorp LLC.

Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Description and explanatory models. *Journal of Early Adolescence*, *37*(1), 85-128.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Thomas, A. (2005). OpenBUGS (Third ed.) [Computer software manual]. OpenBUGS Foundation.

Thomson, G. H. (1919). A direct deduction of the constant process used in the method of right and wrong. *Psychological Review*, *26*, 454-466.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*(38), 406-427.

Thurstone, L. L. (1935). *The vectors of mind.* The University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis: a development and expansion of the vactors of the mind.* The University of Chicago Press, Chicago, USA.

Ueda, N., & Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, *11*, 271-282.

van der Laken, P. (2017, September). *Simpson's paradox: Two hr examples* *with r code*. Blog. Retrieved from `https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/`

van der Linden, W. J. (Ed.). (2016a). *Handbook of item response theory, volume 1 - models*. Taylor & Francis Group.

van der Linden, W. J. (Ed.). (2016b). *Handbook of item response theory, volume 2 - statistical tools*. Taylor & Francis Group.

van der Linden, W. J. (Ed.). (2016c). *Handbook of item response theory, volume 3 - applications*. Taylor & Francis Group.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer-Verlag New York.

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric and nonparametric hierarchical models. *Statistica Neerlandica*, *58*(2), 220-233.

Zheng, X., & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with GLLAMM. *The Stata Journal*, *7*(3), 313-333.