Friedrich-Schiller-Universität Jena
Fakultät für Sozial- und Verhaltenswissenschaften
Institut für Psychologie

Dissertation

Item Nonresponses in Educational and Psychological Measurement

Dissertation
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena
von Dipl.-Psych. Norman Rose
geboren am 16.04.1974 in Jena

Gutachter:

1. Prof. Dr. Rolf Steyer (Friedrich-Schiller-Universität Jena)

2. Prof. Dr. Benjamin Nagengast (Eberhard-Karls-Universität Tübingen)

3. Dr. Matthias von Davier (Educational Testing Service Princeton)

Tag der mündlichen Prüfung: 18. Februar 2013

 ometimes you eat the bear and sometimes the bear, well, he eats you.

The Stranger in Coen brothers' movie "Big Lebowski" (1998)

Dedicated to

my parents Ludmilla & Helmut Rose

and

Klaus-Jürgen Günther

Acknowledgements

I would like to thank my supervisor Prof. Dr. Rolf Steyer who influenced my methodological thinking fundamentally. I am truly indebted and thankful to my second supervisor Dr. Matthias von Davier who gave me confidence and continuous encouragement to carry on with the research that eventually led to this thesis. I am deeply grateful to Prof. Dr. Benjamin Nagengast for his advice and support.

This dissertation would not have been possible without the love, encouragement and support of my parents Ludmilla B. Rose and Helmut K. Rose. Their compassion and understanding kept me writing in difficult times. It may seem paradox, but reminding me that a doctoral degree and a professional career is not everything in life gave me the strength to finish this thesis. I am deeply grateful to Klaus-Jürgen Günther who has supported me for many years in every imaginable way.

I am truly indebted and thankful to Christiane Fiege and Tim Loßnitzer who literally went with me as colleagues and as friends, to continue on the stony road to the doctoral degree. I owe sincere and earnest thankfulness to Katrin Schaller who helped me through the jungle of bureaucracy. Marcel Bauer was exceptionally supportive at all times regarding any type of IT matter I was faced with. Suggestions given by Marlena Itz and Anna-Lena Dicke have been a great help in improving my English skills. I am obliged to many of my colleagues such as Axel Mayer, Prof. Dr. Felix Thoemmes, Anja Vetterlein and Erik Sengewald. My special thanks are extended to Prof. Dr. Ulrich Trautwein and my colleagues at the Center for Educational Science and Psychology in Tübingen who gave me the opportunity to finish this work.

I am particularly grateful to my friends Andre Güttler, Antje Thomas, Dr. Hendryk Böhme, Sebastian Born, Thomas Höhne and Dennis Kiessling who have been my companions for many years.

Finally, my heartfelt thanks go to Jessika Golle for being with me on the wonderful and mysterious journey of life.

Norman Rose
Tübingen, December 2012

Zusammenfassung

Fehlende Werte (*missing data*) sind in der psychologischen und empirischen Bildungsforschung ein ubiquitäres Problem. Seit Jahrzehnten herrscht eine kontroverse Diskussion um die Frage, wie fehlende Werte in der psychologischen Diagnostik und der Leistungsdiagnostik adäquat zu berücksichtigen sind. Selbst in renommierten internationalen Forschungsprogrammen und Large Scale Assessments wie z. B. PISA (Program for International Student Assessment), TIMSS (Third International Mathematics and Science Study) oder PIRLS (Progress in International Reading Literacy Study) konnte bislang keine allgemein akzeptierte Methodologie zur Berücksichtigung fehlender Werte etabliert werden. Seit Ende der 90iger Jahre des letzten Jahrhunderts sind im Rahmen der Item Response Theorie multidimensionale Modelle für fehlende Daten entwickelt worden. Diese weisen jedoch den Nachteil einer hohen Modellkomplexität auf und beruhen zudem auf Annahmen, die bisher kaum Gegenstand des wissenschaftlichen Diskurses waren. Betrachtet man die Problematik fehlender Werte formal auf der Basis statistischer Theorien, so ist die korrekte Behandlung fehlender Werte indiziert, um die Effizienz der Parameterschätzungen zu steigern sowie Schätzfehler zu vermeiden. Ergebnisse empirischer Untersuchungen weisen jedoch darauf hin, dass IRT-basierte Item- und Personenparameterschätzer recht robust gegen fehlende Werte sind. Solche Befunde stellen den Nutzen von komplexen IRT-Modellen für fehlende Werte zunächst in Frage.

Die vorliegende Arbeit besteht aus zwei Teilen. Nach Einführung der Theorie fehlender Daten im Kontext der Testtheorie, wird im ersten Teil der Einfluss fehlender Werte auf verschiedene Item- und Personenparameterschätzer in Messmodellen für dichtome Items untersucht. Im zweiten Teil werden bestehende Ansätze zur Behandlung fehlender Werte in Messmodellen untersucht und weiterentwickelt. Der Fokus dieser Arbeit liegt auf systematisch fehlenden Item-Antworten (*nonignorable missing data*). Die verschiedenen Ansätze werden kritisch verglichen und Empfehlungen für die Anwendung gegeben.

Der Einfluss fehlender Werte auf verschiedene Item- und Personenparameterschätzer wurde sowohl analytisch untersucht als auch empirisch unter Verwendung von simulierten Daten demonstriert. Für systematisch fehlende Werte liesen sich deutliche Schätzfehler für Personen- und Itemparameterschätzungen in IRT-basierten Messmodellen nachweisen.

Diese Ergebnisse unterstreichen den Bedarf geeigneter Methoden zur Berücksichtigung fehlender Item-Antworten. Es wurde gezeigt, dass einfache ad-hoc Methoden - wie bspw. die Kodierung fehlender Werte als falsche Antworten oder als teilweise gelöst - theoretisch nicht zu rechtfertigen sind und zudem die Testfairness sowie die Validität der Testergebnisse gefährden. Ein weiterer Ansatz zur Behandlung fehlender Werte stellt das Nominal Response Modell (NRM) für fehlende Item-Antworten dar, bei dem das Fehlen einer Item-Antwort als zusätzlich Antwortkategorie betrachtet wird. Die Wahrscheinlichkeit fehlender Daten wird somit explizit modelliert, wodurch der Fehler in den Item- und Personenparameterschätzern korrigiert werden soll. Es konnte jedoch analytisch gezeigt werden, dass das NRM auf starken Annahmen beruht und seine Anwendung somit auf wenige Anwendungsfälle beschränkt ist.

Multidimensionale IRT-Modelle (MIRT-Modelle) für fehlende Daten gehören zu den modernen modellbasierten Ansätzen zur Behandlung fehlender Werte. Die theoretische Fundierung dieser Modelle wurde detailliert dargestellt. Es konnte gezeigt werden, dass MIRT-Modelle für fehlende Item-Antworten Spezialfälle von *selection models* und *pattern mixture models* für systematisch fehlende Werte in Modellen für latente Variablen sind. Es sind in den vergangenen Jahren verschiedene MIRT-Modelle für fehlende Werte in der Literatur beschrieben worden, die zumeist als äquivalent betrachtet werden. Zwei Klassen von Modellen können dabei unterschieden werden: *between-item-* und *within-item* multidimensionale Modelle. In der vorliegenden Arbeit konnte gezeigt werden, dass diese Modelle nicht per se äquivalent sind. Die Frage der Äquivalenz von Modellen wird im wissenschaftlichen Diskurs zumeist hinsichtlich des Kriteriums der Modellpassung diskutiert. In Modellen zur Behandlung fehlender Werte ist dieses Kriterium jedoch nicht hinreichend. Die Konstruktion der latenten Variable, die von theoretischem Interesse ist, sowie die Reduktion des Schätzfehlers aufgrund fehlender Werte müssen ebenfalls berücksichtigt werden, um konkurrierende Modelle für fehlende Werte hinsichtlich ihrer Äquivalenz beurteilen zu können. Es wird weiterhin ein allgemeines Rahmenkonzept für MIRT-Modelle vorgeschlagen, in dem verschiedene between- und within-item multidimensionale IRT-Modelle verortet und hinsichtlich der verschiedenen Aspekte der Modelläquivalenz beurteilt werden können. Aufgrund ihrer einfachen Spezifizierbarkeit und Interpretierbarkeit werden between-item multidimensional IRT Modelle für die Praxis empfohlen.

Die Modellkomplexität der MIRT-Modelle für fehlende Item-Antworten hängt wesentlich von der Zahl der Items und der latenten Variablen im Modell ab. Für die stochastische Modellierung des Fehlens von Werten verdoppelt sich nicht nur die Zahl der mani-

festen Variablen sondern auch die Anzahl latenter Variablen im Messmodell nimmt zu. Neben den latenten Dimensionen, die von theoretischem Interesse sind, wird eine latente Antworttendenz (*latent response propensity*) eingeführt. In den meisten Anwendungen wird angenommen, dass diese latente Antworttendenz eine eindimensionale Variable ist. Dies ist jedoch eine sehr starke und oft ungeprüfte Annahme. Die Ergebnisse dieser Arbeit zeigen, dass MIRT-Modelle den Schätzfehler nicht oder nur unzureichend korrigieren, wenn die Dimensionalität der latenten Antworttendenz nicht korrekt berücksichtigt wird. Leider sind hochdimensionale IRT-Modelle noch immer eine numerische Herausforderung. Aus diesem Grund werden latenten Regressionsmodelle und Mehrgruppen-IRT Modelle für fehlende Item-Antworten als sparsamere Alternativen zu MIRT-Modellen dargestellt. Die Verbindung zwischen den verschiedenen Modellansätzen wird ausführlich erläutert und die jeweils zugrunde liegenden Annahmen werden diskutiert.

Abschließend konnte gezeigt werden, dass fehlende Werte aufgrund von Auslassungen (*omitted items*) während des Tests im Vergleich zu fehlenden Item-Antworten am Ende des Tests (bspw. aufgrund von Zeitmangel; *not-reached items*) unterschiedliche stochastische Eigenschaften aufweisen. Diese Unterschiede haben Implikationen hinsichtlich der Behandlung der fehlenden Werte. Während Auslassungen durch MIRT-Modelle adäquat berücksichtigt werden können, sind nicht erreichte Items am Testende durch Regressionsmodelle oder Mehrgruppen-IRT Modelle zu berücksichtigen. Da fehlende Werte aufgrund ausgelassener und nicht erreichter Items häufig gemeinsam auftreten wurde ein Modell zur simultanen Modellierung beider Formen fehlender Werte abgeleitet. In einer abschließenden Diskussion werden die Ergebnisse zusammengefasst, Einschränkungen der verschiedenen Ansätze kritisch diskutiert und Empfehlungen für die Anwendung gegeben. Bestehende Forschungsfragen und bislang ungelöste Probleme werden diskutiert.

Abstract

The question of how to handle missing responses in psychological and educational measurement has been repeatedly and controversially debated for decades. Even in highly respected international studies and large scale assessments, such as the PISA (Program for International Student Assessment), TIMSS (Third International Mathematics and Science Study), and PIRLS (Progress in International Reading Literacy Study) a generally accepted methodology for missing data is still lacking. Since the late 1990s multidimensional item response theory (MIRT) models for item nonresponses have been developed. These models become quickly complex in application and rest upon assumptions that are usually not critically addressed. Although statistical theory of missing data suggests adequate handling of missing responses to avoid inefficient and biased parameter estimation, there is empirical evidence that IRT-based parameter estimation is fairly robust against missing responses. That may question the need for sophisticated IRT model-based approaches. For that reason this thesis consists of two major parts. After the introduction of the missing data theory in the context of educational and psychological measurement, the impact of item nonresponses to item- and person parameter estimates are examined in the first part. In the second part existing approaches to handle missing responses are scrutinized and further developed. The different methods are critically compared and recommendations will be given as to which approaches are appropriate. The considerations are confined to dichotomous items that are still common in many tests and assessments.

The impact of missing responses to item and person parameter estimates was shown analytically and empirically using simulated data. The results show clearly that ignoring systematic missing data leads to biased item and person parameter estimates in IRT models. The findings highlight the need for appropriate methods to handle item nonresponses properly. It could be shown that simple ad-hoc methods such as incorrect answer substitution (IAS) or partially correct scoring (PCS) are not justifiable theoretically and threaten the test fairness and the validity of test results. The nominal response model (NRM) for item nonresponses is an alternative approach that was examined. In this model item nonresponse are regarded as an additional response category. However, the NRM rests upon strong assumptions and, therefore, its applicability is limited.

MIRT models for missing responses rank among modern model-based approaches. The underlying rationale of these models is outlined in detail. It could be shown that MIRT models for item nonresponses are special cases of selection models and pattern mixture models for latent trait models with particular assumptions. Different MIRT models are discussed in the literature and are typically regarded to be equivalent. Two classes of MIRT models can be distinguished: between- and within-item multidimensional IRT models. In this thesis it is shown that these models are not equivalent per se. Typically, the question of model equivalence is considered with respect to the model-fit. In models for item nonresponses the criterion of model-fit is insufficient to judge equivalence of alternative models. The equivalence in the construction of the latent variable of interest and the bias reduction are additional criteria that need to be considered. A common framework of IRT models for item nonresponses is presented. Different between- and within-item multidimensional IRT models are rationally developed, taking the issue of model equivalence into account. Between-item multidimensional models are easy to specify and to interpret and are recommended as the models of choice.

The disadvantage of MIRT models for item nonresponses is their complexity. Besides the latent variables of theoretical interest, a latent response propensity is introduced to model the missing data mechanism. Typically, unidimensionality of the latent response propensity is assumed in application. This is a strong and often untested assumption. It could be demonstrated that MIRT models fail to correct for missing data if multidimensionality of the latent response propensity is not taken into account. Hence, the number of manifest and latent variables can become fairly large in MIRT models for item nonresponses. Unfortunately, high-dimensional MIRT models are still computationally challenging. For that reason more parsimonious and less demanding latent regression IRT models and multiple group IRT models are derived as an alternative. The relationship between these models and the MIRT models is demonstrated. Finally, it is shown that missing responses due to omitted and not-reached items have different properties suggesting different treatments of them in IRT measurement models. Whereas omitted responses can be appropriately handled by MIRT models, not-reached items need to be taken into account by latent regression models. Since real data sets typically suffer from both, omitted and not-reached items, a joint model is introduced that account for both types of missing responses. The thesis ends with a final discussion in which the findings are summarized and recommendations for real applications are given. Unsolved problems and remaining research questions are outlined.

Contents

List of Figures

List of Tables

List of Listings

Acronyms and Symbols

Variables

Y_i	Manifest variables that are part of the measurement model of ξ
Z_j	Covariates as gender, social status, country etc.
D_i	Missing data indicator variable for Y_i
ξ	Latent person variable that is intended to measure
$\hat{\xi}$	Estimator of the latent person variable ξ
θ	Latent response propensity
$\hat{\theta}$	Estimator of the latent response propensity θ

Indices

I	Number of manifest variables Y_i $(i = 1, \ldots, I)$
J	Number of manifest variables Z_j $(j = 1, \ldots, J)$

Symbols & Parameters

α_i	Item discrimination of item Y_i
β_i	Item difficulty of item Y_i
$\boldsymbol{\iota}$	Vector of parameters of a particular measurement model of $\boldsymbol{Y}$
$\boldsymbol{\phi}$	Index of the missing data mechanism
$\boldsymbol{\upsilon}$	The joint parameter vector $\boldsymbol{\upsilon} = (\boldsymbol{\iota}, \boldsymbol{\phi})$
$\hat{\boldsymbol{\iota}}_{ML}$	Vector valued maximum likelihood estimator of $\boldsymbol{\iota}$
$\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$	Short form of the likelihood function $\mathcal{L}(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\iota})$
$\ell(\boldsymbol{y}; \boldsymbol{\iota})$	Short form of the log-likelihood function $\ell(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\iota})$
$g(.)$	Density function
$g(. \mid .)$	Conditional density function
$f(.)$	General function
$. \perp .$	Stochastic independence
$. \perp . \mid .$	Conditional stochastic independence
$. \not\perp .$	Stochastic dependency

$.\not\perp.\,\vert\,.$	Conditional stochastic dependency
$.\vdash.$	Regressive independence
$.\vdash.\,\vert\,.$	Conditional regressive independence
$.\nvdash.$	Regressive dependency
$.\nvdash.\,\vert\,.$	Conditional regressive dependency

Acronyms and Abbreviations

CTT	Classical test theory
EAP	Expected a posteriori estimator
FIML	Full information maximum likelihood
IRT	Item response theory
LRM	Latent regression model
MAP	Maximum A Posteriori
MAT	Multidimensional Adaptive Testing
MIRT	Multidimensional item response theory
MG-IRT	Multiple group IRT
ML	Maximum Likelihood
MLE	Maximum Likelihood estimator
MMLE	Marginal maximum likelihood estimation
NAEP	National Assessment of Educational Progress
SLM	Selection models
PMM	Pattern Mixture Models
IAS	Incorrect answer substitution for missing data
ICC	Item characteristic curve
PCS	Partially correct scoring of item nonresponses
w.r.t.	with respect to

1 Introduction

Missing data are an ubiquitous problem in most empirical sciences. Unfortunately, most of the commonly used statistical models and their implementations in available software do not allow for missing data (Schafer & Graham, 2002). Cases with incomplete data are commonly excluded from the analyses. Although standard methods for data analysis become applicable by deletion of incomplete data records (listwise deletion), valuable information is wasted and the problem of missing data is even aggravated. Correct statistical inference using listwise deletion rest on strong assumptions. A loss of efficiency and potentially biased parameter estimates result. Enders stated that researchers slowly adopt appropriate missing data methods (Enders, 2010) developed in the last decades. Nevertheless, missing data methods are becoming increasingly available in modern statistical software. Furthermore, many textbooks give non-technical introductions to the issue of missing data and the approaches to handle them (Enders, 2010; McKnight, McKnight, Sidani, & Figuerdo, 2007). For that reason missing data are more and more taken into account in real applications. However, missing item responses in measurement models are much less addressed and there seems to be considerable disagreement between researchers on how to handle them. This work is intended to examine widely used methods to handle missing responses in dichotomous items of measurement models used in psychological and educational assessments. The different methods are scrutinized with respect to their appropriateness considering the underlying, and often implicit, assumptions. There is a major focus on systematic item nonresponses that result in biased item and person parameter estimates. Such nonignorable missing responses can be handled in advanced model-based methods, such as multidimensional Item Response Theory (IRT) models for missing data (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko, Glas, Bosker, & Luyten, 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose, von Davier, & Xu, 2010). These models are extended and further developed. Unfortunately, high-dimensional IRT models are still numerically and computationally challenging (Cai, 2010). Therefore, alternative less demanding models are derived. A common framework for all these models will be introduced and the different modelling approaches will be critically compared.

Not only missing responses in itself but also their improper handling threatens reliability of the results as well as test fairness and validity of the test results. This will be demonstrated by examining traditional methods for item nonresponses critically. The development of alternative methods for item nonresponses is vital in order to obtain reliable results in psychological and educational testings. This is essential if test results serve as a basis for far-reaching decisions. For example, in international educational effectiveness studies, such as Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Studies (TIMSS), the National Assessment of Educational Progress (NAEP), and other large scale assessments are used to quantify students achievement. These studies are typically low-stakes assessments with significant proportions of missing responses. Appropriate methods are required to ensure comparability of test results and facilitate reliability of results.

Missing item responses in psychological and educational measurement In 1974 Lord published a paper entitled „Estimation of latent ability and item parameters when there are omitted responses“. He stated that simply to score omitted items as incorrect answers is not appropriate. 37 Years later, in April 2011, Culbertson presented a speech at the Annual Meeting of the National Council on Measurement in Education in New Orleans entitled „Is It Wrong? Handling Missing Responses in IRT“. In this talk he concluded „... the most pragmatic choice in high-stakes testing may be to continue to treat omitted responses as incorrect and to encourage examinees to respond to all items.“ In fact the question of how to treat missing responses has remained unanswered for more than four decades. The best way to handle missing data is to prevent them from occurring by appropriate study designs, measurement instruments, and instructions (e. g. McKnight et al., 2007). Especially in low-stakes assessments, however, an excellent study design and even the best instruction will hardly prevent the occurrence of missing data. In fact, complete data sets are extremely rare in empirical studies. Missing data occur for a variety of reasons. They are more likely in some research areas than in others. For example, they are more likely in surveys that focus on private issues such as relationship problems, money matters or health problems. Questions or items addressing those issues might be regarded as offensive or improper. Study participants may feel insecure about the data security, leading to omissions of critical items. Therefore, the willingness to provide information about such issues is typically limited. Apart from the subject of the study, longitudinal studies suffer typically much more from missing data than cross-sectional studies due to attrition or study drop-out (McKnight et al., 2007; Peugh & Enders, Winter 2004). In edu-

cational testings the willingness to respond to all items of an achievement test is typically higher in high stakes than in low-stakes assessments. Therefore, the omission rates are commonly higher in low-stakes assessments. Especially in achievement tests the time to process the items is usually limited resulting in not-reached items. Apart from unplanned missing data due to omissions, not-reached items, or not codable responses, there are planned missing data designs that are applied to lower respondents' burden and to lower costs (Graham, Taylor, Olchowski, & Cumsille, 2006; Graham, 2009). In educational large scale assessments planned missing data are common due to multi-matrix designs, where only a subset out of hundreds of items is presented to each test taker (Frey, Hartig, & Rupp, 2009; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Thomas, Raghunathan, Schenker, Katzoff, & Johnson, 2006). Almost all educational large scale assessments such as PISA, TIMSS, and NAEP used multi-matrix designs. However, depending on the test design and the choice of the model used for data analysis the handling of these kinds of missing data is not per se trivial (Frey et al., 2009). To make matters worse, missing responses in single items usually result from planned missing data as well as from omissions and not reaching the end of the test. Hence, different kinds of missing data can occur simultaneously in a single data set, and perhaps need to be treated differently.

Typology of missing data In general, missing data can be classified in many different ways. Almost all research on missing data since the late 1970s rests upon the taxonomy introduced by Rubin (1976). He distinguishes three different missing data mechanisms; (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (NMAR). These kinds of missingness will be introduced in detail in Chapter 2. In very simple terms, MCAR means that the missingness of variables is independent of the variables considered in the study. If the missing data are MAR, missingness depends exclusively on observable variables. If the missing data are NMAR, missingness depends on unobservable but important variables of interest in the study. Item nonresponses that are MCAR or MAR are called *ignorable*, whereas those that are MNAR are *nonignorable*. As Schafer and Graham (2002) noted, the term missing data mechanism refers neither to a process nor to the cause of missingness. Furthermore, ignorability of the nonresponse mechanism does typically not mean that missing responses can be neglected in data analyses. In fact most of modern missing data handling methods are appropriate to account for ignorable instead of nonignorable missing data. Despite the disadvantage of a somewhat confusing terminology used in Rubin's framework, its great value has been undoubtedly proven. A major advantage of his approach is that theory directly implies

how to handle missing data appropriately. In fact, Rubin introduced not simply a typology of missing data, but rather he proposed a framework of inference from incomplete data (Schafer & Graham, 2002).

However, recently McKnight et al. (2007) proposed an alternative classification scheme that is completely compatible with Rubin's taxonomy. Referring to Cattell's data box (1966), these authors stated that three facets can be considered to describe missing data: *individuals*, *variables*, and *occasions*. Missing data can occur with respect to each of these facets. For instance, subgroups of individuals that refuse to participate in a study or are not available can lead to missing data. This is called *unit nonresponse*. If participants and nonparticipants differ systematically with respect to variables of interest, unit-nonresponses result in unrepresentative samples, which threatens the generalizability of results. If some items are refused to be answered, not-reached or not appropriately answered (not codeable) the resulting missing data are denoted as *item-nonresponses*. Finally, *wave-nonresponses* can result in longitudinal studies when participants are not available at each measurement occasion. Many of these studies suffer from drop out of participants over the course of the study. The differentiation of nonresponses in unit-, item-, and wave-nonresponses does not contradict the missing data mechanisms introduced by Rubin. In fact, each kind of nonresponses - unit-, item-, and wave-nonresponses - can be MCAR, MAR, or MNAR.

Methods for missing data In data sets unit-nonresponse, item-, and wave-nonresponses differ with regard to their appearance. Wave-nonresponses due to drop out over the study period result in typical monotone missing data pattern with increasing proportions of nonresponses over time(e. g. Little & Rubin, 2002; McKnight et al., 2007). Item-nonresponses result in general non-monotone missing data patterns. The problem of unit-nonresponses can even be present in seemingly complete data sets if study participants in a survey respond to all items but are not representative with respect to the population of interest. Due to the peculiarities missing data handling methods for unit-, item-, and wave-nonresponses differ. Most of the common statistical models and their estimation algorithms are developed for complete data, which cannot properly account for missing responses. That is why inefficient missing data handling methods as listwise or pairwise deletion are still widely used (Allison, 2001; Enders, 2010; Schafer & Graham, 2002). Apart from traditional methods such as listwise and pairwise deletion, mean-imputation, etc., modern methods have been developed that can be divided into (a) weighting approaches, (b) imputation based methods, and (c) model-based approaches (Lüdtke, Rob-

itzsch, Trautwein, & Köller, 2007). Weighting methods are most appropriate to account for unit-nonresponses (L. Li, Shen, Li, & Robins, 2011; Little, 1988a). In unrepresentative samples, cases are weighted in a way that the sampling of a representative sample is emulated (e. g. inverse probability weighting; IPW). Item- and wave-nonresponses can be better handled by imputation- and model-based methods. Imputation based methods rest upon replacing missing data by expected, predicted, or plausible values. The filled-in data sets are analyzed in a subsequent step by means of standard complete-data methods. Among different imputation techniques, multiple imputation (MI) has become the method of choice (Rubin, 1987, 1996; Schafer, 1997). Accordingly, most of the commonly used software packages in social and behavioural sciences provide MI for continuous and categorical variables (Enders, 2010). Model-based methods directly account for nonresponses in the stage of parameter estimation. Hence, a preceding data augmentation is not required. One of the most popular model-based approaches is Full Information Maximum Likelihood (FIML) estimation (Arbuckle, 1996; Enders, 2001b), which is conceptually close to multiple group (MG) approaches for missing data (Muthén, Kaplan, & Hollis, 1987). The Expectation Maximization (EM) algorithm is an alternative methods to obtain unbiased ML estimates in presence of missing data (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 2008). However, FIML, EM, MG-models, and MI require that the missing data mechanism is MCAR or MAR. Only a few approaches exist to tackle the problem of nonignorable missing data. Heckman (1976, 1979) proposed selection models (SLM) to handle nonignorable missing data in normally distributed variables. This approach was extended to regression models with non-normal variables (Dubin & Rivers, 1989). Pattern mixture models (PMM) are an alternative class of models for missing that are NMAR (Glynn, Laird, & Rubin, 1986; Little, 1993, 2008; Little & Rubin, 2002). Both SLMs and PMMs are not frequently used (Enders, 2010). Whereas SLMs rely on very strong distributional assumptions PMMs are not identified without restrictions on unknown parameters. This might prevent popularity of models for nonignorable missing data, although the ignorability assumption is hardly tenable in many applications.

Missing data in latent variable models To account for unreliability as well as situational effects, method effects, and other sources of variances in test data, the use of measurement models is common in psychological and educational measurement (e. g. Bollen, 1989; Brennan, 2006; Eid & Diener, 2006; Embretson & Reise, 2000; Rost, 2004; Steyer, 1989; Steyer, Schmitt, & Eid, 1999; Steyer & Eid, 2001; Steyer, 2001). In measurement models of Classical Test Theory (CTT) and Item Response Theory (IRT) latent variables

are introduced. The individual values of these variables represent persons' individual trait level, free of measurement error. Latent variable models additionally include parameters that describe the relationship between latent and manifest variables such as factor loadings, measurement intercepts, item difficulties, and item discriminations. In IRT models such parameters are called item parameters in contrast to person parameters, which are individual values of the latent variables. In most psychological and educational testings both item and person parameters are aimed to be inferred from observed data. Unfortunately, sample based estimates of item and person parameters can be seriously biased due to missing data (e. g. Culbertson, 2011, April; de Ayala, Plake, & Impara, 2001; Glas, 2006; Rose et al., 2010). Given the nonresponse mechanism is MCAR or MAR, FIML can be used for SEM with latent variables (Arbuckle, 1996; Enders & Bandalos, 2001). In IRT, joint maximum likelihood (JML) estimation and marginal maximum likelihood (MML) estimation can also be seen as full information ML estimation techniques, since each observed response is included. Hence, both estimation techniques are reliable given the missing data mechanism is ignorable. Sometimes the MAR assumption is only tenable if observable covariates are included that are related to missingness but which are not part of the target model. The inclusion of such auxiliary variables without changing the meaning of parameters of the target model is not trivial (e. g. Allison, 2003; de la Torre, 2009; Graham, 2003). Alternatively, data augmentation methods as MI can be used. With the introduction of multiple imputations by fully conditional specifications (T. Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; Van Buuren, 2007), MI has become an alternative method to handle item-nonresponses in dichotomous and other categorical items (2010). However, as previously noted, these methods require that the missing data mechanism is ignorable. Methods and models for nonignorable missing responses are of great interest in psychological and educational measurement. There is strong evidence that item nonresponses due to omitted or not-reached items in tests are related to unobserved person characteristics of test takers. For example, Rose, von Davier and Xu (2010) showed that in PISA 2006 data the proportion of correctly answered items is substantially correlated with the proportion of missing responses. Test takers with lower proportions of correctly answered items had on average more missing responses. Similarly, Culbertson (2011, April) found in high-stakes educational assessments that the probability of omissions of items with short and open response formats increases with decreasing proficiency levels. These findings indicate a relationship between missingness and test performance, which suggests a stochastic dependency between the occurrence of item nonresponses and persons' proficiency. Hence, the missing data depends on un-

observable variables, which is distinctive for a nonignorable missing data mechanism. Korobko, Glas, Bosker, and Luyten (2008) found that self selection in examination subjects and, therefore, in achievement tests result in nonignorable missing data as well. Despite today's heightened awareness of the problem of missing data and the ongoing development of elaborate missing data methods, ad hoc methods such as scoring missing responses as wrong or partially correct are still common in many educational assessments (Culbertson, 2011, April; Rose et al., 2010). Interestingly, such questionable methods are typically applied with full knowledge of the potentially detrimental effects due to external pressures (e. g. Culbertson, 2011, April). This fact underlines the need for appropriate methods for handling nonignorable item nonresponses in psychological and educational measurement.

In recent years, IRT models for nonigorable missing data were introduced (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). Multidimensional IRT (MIRT) models and nominal response models (NRM;Moustaki & O'Muircheartaigh, 2000) for item-nonresponses can be distinguished. In 2010 Rose et al. proposed latent regression models and multiple group IRT models for nonignorable missing data in measurement models. The same authors found that IRT based item and person parameter estimates seemed to be quite robust to nonignorable missing responses if they are simply ignored. This raises the question whether elaborate model-based approaches are really needed. In fact, it was repeatedly found that incorrect answer substitution (IAS), partially correct scoring (PCS) and other ad hoc methods results in even more biased parameter estimates than ignoring missing responses even if they are NMAR.

Outlook This work focuses on nonignorable item nonresponses in IRT measurement models for dichotomous items. In Chapter 2 Rubin's typology of missing data will be introduced and adapted to the context of psychological and educational measurement. In a first step the different nonresponse mechanisms will be defined with respect to single items. Since the missing data mechanisms can vary across items in a single test, the same nonresponses mechanisms will be defined next with respect to the whole response vector in a second step. In IRT measurement models the manifest items constitute a measurement model of latent variables. What are the implications of the different missing data mechanisms with respect to true score variables and the latent ability variable constructed in the model? This will be studied analytically in the last part of Chapter 2.

In Chapter 3, the impact of item nonresponses to sample based item and person pa-

rameter estimates will be studied. The considerations will be confined to the one- and two parametric IRT models. Hence, the bias of estimated item difficulties and item discriminations are studied. In IRT different person parameters exist. Here the bias of maximum likelihood (ML) estimates, weighted maximum likelihood (WML) estimates, and expected a posteriori (EAP) estimates are studied. Unfortunately, the bias of IRT based parameter estimates are difficult to study by analytical means. For that reason, the impact of missing responses will also be studied priorly with respect to sum score and the proportion correct score. Most tests constructed based on CTT uses the sum score or functions of it as person parameter estimates. Expected values of items are commonly used as population specific measures of item difficulty. Here the bias of the sum score, the proportion correct score and the item means are examined. The reason is that the impact of missing data to CTT based item and person parameter estimates can easily be studied analytically. The results serve to generate hypotheses about biasedness of IRT based parameter estimates, which will be verified by means of data simulations. The demand for appropriate models that account for item-nonresponses will be justified. Finally, it will be studied how missing data affect measures of accuracy of person parameter estimates, such as standard errors and the marginal reliability. The results of the study of biasedness of item and person parameter estimates in Chapter 3 motivate the detailed examination and the further development of IRT based methods for item nonresponses.

In Chapter 4 a short general introduction to existing missing data methods will be given. Subsequently, traditional ad hoc methods such as incorrect answer substitution (IAS) and partially correct scoring (PCS) will be examined with respect to their suitability to handle missing responses. In the terminology of missing data theory, these methods are imputation based approaches. Accordingly, IAS and PCS will be studied with respect to the theoretical assumptions underlying the respective imputation model. In a further step, model-based approaches will be studied starting with the nominal response model for nonignorable missing data (Moustaki & Knott, 2000; Moustaki & O'Muircheartaigh, 2000). It will be clarified under which conditions the NRM can be used to account for nonignorable missing responses. This work primarily focuses on MIRT models for nonignorable missing data (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). The rational of these models will be outlined. The essential idea is the introduction of a latent response propensity as a function of the persons that determine the specific item response propensities. The underlying assumptions of these MIRT models and their implications will be discussed critically. Based on these considerations, the range of MIRT models will be ex-

tended to allow for less restrictive and more flexible MIRT models for missing responses. In the existing literature between- (B-MIRT) and within item-multidimensional IRT (W-MIRT) models for item nonresponses can be distinguished. The two classes of models are typically considered to be equivalent (e. g. Holman & Glas, 2005; Rose et al., 2010). In fact, Rose et al. (2010) could show that in one-parametric models B- and W-MIRT models are equivalent. As it will be demonstrated here, this is not necessarily true for two-parameter models. Furthermore, the interpretation of item and person parameters in W-MIRT models is different compared to B-MIRT models. Here, two different W-MIRT models will be rationally derived starting with the definition of latent variables in the model. The implications of the definitions with respect to item parameters will be used to derive constraints, required to specify the different models in existing software. The question of which model should be preferred in applications will be discussed. The resulting models are critically compared considering the issue of model equivalence. It will be shown that equivalent model fit is not sufficient to regard two MIRT models for item nonresponses as being equivalent. Unfortunately, MIRT models for nonignorable missing data are very complex. With IRT models including latent regression models (LRM) and multiple group (MG) IRT models, simpler models have been proposed (Rose et al., 2010). The relation between these models and the more complex MIRT models will be explained. Advantages and disadvantages of each model are discussed and a common framework of IRT model-based approaches to handle nonignorable item-nonresponses is proposed.

Should omitted and not-reached items be treated differently? This question will be answered in Section 4.5.6 considering the assumption underlying MIRT models for missing data. A joint model for omitted and not-reached items will be introduced. Finally, in a general discussion (Chapter 5), the findings of this work will be summarized and recommendations for applied researchers will be given. Unanswered questions are discussed to facilitate future research in this area.

2 Theory

In 1976 Rubin provided a first comprehensive typology of missing data based on the so-called missing data mechanism (Little & Rubin, 2002; Rubin, 1976). As Schafer (2002) noted this taxonomy is widely used but less widely understood. Several different factors may contribute to this confusion. Most applied statisticians who deal with the topic of missing data including Rubin do not clearly distinct between realized data and random variables. On the one hand these authors emphasize that the term missing data mechanism refers to the distribution of missingness, i. e. unconditional or conditional probabilities of missing patterns (e. g. Schafer & Graham, 2002). Hence, the used terminology seems to imply that Rubins definitions are based on stochastic dependencies between random variables. However, Rubin (2009) noted: „These definitions [of the missing data mechanisms] are not simply jargon for conditional independence.“It is important to note that Rubin is a Bayesian statistician. Scrutinizing his literature reveals that his definitions are actually based on *posterior* distributions of unobserved values given a particular missing pattern in a concrete sample and given the observed values (Gelman, 2002; Rubin, 2009). This is rarely made explicitly. As common in Bayesian tradition, a distribution of unknown quantities such as parameters *and* unobserved realizations of random variables result from uncertainty given the current state of knowledge. This might explain why missing data literature of Rubin and his colleagues might be somewhat confusing to non-Bayesian statisticians and frequentists. To fill the gap Kenward and Molenberghs (1998) discussed missing data mechanisms in frequentist's terminology under common likelihood inference. In the remainder of this work, likelihood inference as well as maximum likelihood estimation theory will be of major importance. In order to yield a stringent and consistent terminology, Rubin's classification scheme is therefore adapted following Kenward and Molenberghs (1998). The missing data mechanisms will be defined based on stochastic relationships between random variables considered in a particular random experiment. It is important to note that the definitions of MCAR, MAR and NMAR as introduced in this section are consistent with Rubin's taxonomy. That is, data of sufficiently large samples drawn under a specific missing data mechanism as defined below will have the properties that are implied by Rubin's definitions.

In psychological and educational assessments manifest variables often can be classified into items Y_i that constitute the measurement model of a latent variable, and covariates Z_j as background and context variables, which are not part of considered measurement model. Due to this distinction, five missing data mechanisms can be distinguished instead of three; including three different MAR conditions. In a first step, the missing data mechanisms will be defined with respect to single items Y_i and subsequently with respect to the complete response vector $\boldsymbol{Y}$. Throughout this work, it is assumed that the covariates do not suffer from missingness. Hence, in application all values of covariates are fully observed. Before the typology of nonresponse mechanisms is introduced, the random variables and the underlying random experiment are introduced formally. In the final section of this Chapter, the implications of the different kinds of missing data with respect to the distribution of latent variables in the measurement model will be examined.

2.1 Classification of missing data

The most simple case of missing data concerns only a single variable Y. There can be many covariates constituting the multidimensional covariate $\boldsymbol{Z} = Z_1, \ldots, Z_J$. In order to define the missing data mechanism with respect to Y, a response indicator variable D needs to be introduced that indicates the observational status of Y. All variables are random variables on the same probability space $(\Omega, \mathfrak{A}, P)$, with Ω the set of possible outcomes given by

$$\Omega = \Omega_{Z_1} \times \ldots \times \Omega_{Z_j} \times \ldots \times \Omega_{Z_J} \times \Omega_D \times \Omega_Y. \tag{2.1}$$

$\mathfrak{A}$ is a σ-algebra with the set of possible events and P is the probability measure on $\mathfrak{A}$. Hence, Y, D, and $\boldsymbol{Z}$ are random variables in the same probability space, so that $Y\colon \Omega \to \Omega_Y$, $D\colon \Omega \to \Omega_D$, $Z_j\colon \Omega \to \Omega_{Z_j}$, and $\boldsymbol{Z}\colon \Omega \to \Omega_{\boldsymbol{Z}} = \Omega_{Z_1} \times \ldots \times \Omega_{Z_J}$. The covariates Z_j and the variable Y can be continuous or categorical variables. D is the response indicator variable that indicates whether Y is observed or not.

$$D = \begin{cases} 1, \text{if } Y \text{ is observed} \\ 0, \text{if } Y \text{ is not observed} \end{cases} \tag{2.2}$$

Hence, the response indicator variable will be $D = 1$ if the variable Y is observed and $D = 0$ if Y is missing. The probability space $(\Omega, \mathfrak{A}, P)$ refers to the following random experiment: Draw randomly from the multivariate distribution of the random variables Y

and $\boldsymbol{Z}$. Register the values of the covariates $Z_1, \ldots, Z_J$. If Y is observable, register the observed value and assign $D = 1$. If Y cannot be observed then assign $D = 0$.

As previously noted the terms missing data mechanism and distribution or probability of missingness are used interchangeably in missing data literature. The probability of missing Y can be expressed by $P(D = 0)$. In turn, the probability of Y being observed is $P(D = 1)$. The typology of missing data introduced here is based on conditional stochastic (in)dependencies between the missing indicator variable D, the variable Y, and the fully observed covariate $\boldsymbol{Z}$. Hence we can consider the conditional probability $P(D = 1 \,|\, Y, \boldsymbol{Z})$ that Y will be observed given Y and $\boldsymbol{Z}$. Since D is binary, $1 - P(D = 1 \,|\, Y, \boldsymbol{Z}) = P(D = 0 \,|\, Y, \boldsymbol{Z})$ is the probability that Y is not observed *given* Y and $\boldsymbol{Z}$. At first sight it seems counterintuitive to talk about the probability of missing Y given Y. Note, however, that Y and D are random variables with a joint distribution instead of realized data. Accordingly, it is assumed that Y has a conditional distribution $g(Y \,|\, D = 0)$ even if it cannot be observed. For example, if an item was not answered by a particular person it has, nevertheless, a conditional probability of being correctly answered given the person's proficiency level. Hence, we can consider different conditional distributions $g(Y \,|\, D = d, \boldsymbol{Z} = \boldsymbol{Z})$. In particular, we can, at least theoretically, compare the conditional distributions $g(Y \,|\, D = 0, \boldsymbol{Z} = \boldsymbol{Z})$ and $g(Y \,|\, D = 1, \boldsymbol{Z} = \boldsymbol{Z})$ as well as the conditional regressions $E(Y \,|\, D = 0, \boldsymbol{Z})$ and $E(Y \,|\, D = 1, \boldsymbol{Z})$. This fact will be of crucial interest when the implications of the different nonresponse mechanisms will be studied with respect to the latent variables in measurement models (see Section 2.3). So far, it suffices to bear in mind that Y has a distribution given it will be unobserved and that the probability of Y is not being observed given Y is well defined. Based on these considerations, the following missing data mechanisms can be defined.

- The probability of missingness with respect to variable Y is called *missing completely at random* (*MCAR*) if

$$P(D = 1 \,|\, Y, \boldsymbol{Z}) \quad = \quad P(D = 1). \tag{2.3}$$

  That means that the probability of Y being observed is independent from the covariate $\boldsymbol{Z}$ and the considered variable Y.

- The probability of missingness with respect to variable Y is called *missing at ran-*

dom (*MAR*) if the following two conditions are met:

$$(1.) \qquad P(D = 1 \mid Y) \neq P(D = 1) \tag{2.4}$$

$$(2.) \qquad P(D = 1 \mid Y, \boldsymbol{Z}) = P(D = 1 \mid \boldsymbol{Z}) \tag{2.5}$$

Hence, D is stochastically dependent on Y (not MCAR). The probability of observing Y depends on the fully observable covariate $\boldsymbol{Z}$. Given $\boldsymbol{Z}$, however, D is *conditionally* stochastically independent of Y.

- The probability of missingness with respect to variable Y is called *missing not at random* (*MNAR*) if

$$P(D = 1 \mid Y, \boldsymbol{Z}) \quad \neq \quad P(D = 1 \mid \boldsymbol{Z}). \tag{2.6}$$

Hence, D is *not conditionally* stochastically independent of Y given the covariate $\boldsymbol{Z}$. That means that the probability of observing Y depends on Y itself even when controlling for $\boldsymbol{Z}$.

Note that *MAR* does not mean that the occurrence of missing data is purely at random. Quite contrary, there are stochastic dependencies between variables considered and missingness. The term MAR refers to a conditional stochastic independence between variables and their observational status given observable variables. As outlined in the introduction, here the traditionally used terminology will be retained to keep in line with existing literature. Due to the prevalence of Rubin's terminology, it is to be feared that introducing an alternative terminology causes even more confusion. So far, the definitions refer to missingness of a single variable Y. The considered random experiment does not sufficiently reflect the complexity of random experiment that underlies psychological and educational testing. Additional variables need to be considered. The observational units in these assessments are real persons who need to be formally considered by introduction of a person variable. Person characteristics aimed to be measured are conceptualized as theoretical constructs and are constructed as latent variables in appropriate measurement models. Such measurement models typically consist of more than one manifest variable. Hence, the nonresponse mechanisms need not only to be defined with respect to single items but also with respect to the complete response vector. Accordingly, the response indicator variable becomes multidimensional. In the following section the missing data mechanisms are defined such that the complexity of psychological and educational measurement is taken into account.

2.2 Missing Data in the Context of Measurement Theory

Although a comprehensive introduction to psychometrics and measurement theory is far beyond the scope of this work, a few notes on these topics will point to the peculiarities that need be considered when dealing with missing data in psychological and educational measurement. For a comprehensive introduction to test theory and measurement see Borsboom (2005) , Hopkins (1998), McDonald (1999), Moosbrugger and Kelava (2011), Rost (2004), Steyer and Eid (2001), and Thissen (2001).

Typically, in the process of measurement symbols are assigned to persons under study that should represent the particular characteristic of interest. In most applications the symbols are numerical values whose relationships reflect relationships of the characteristics being measured. For instance, the intelligence of persons is expressed by their intelligence quotient, possibly the best known standardized test score. Higher numerical values should indicate higher levels of a person's intelligence. Other well-known examples are tests developed to assess personality traits such as Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, known as the Big Five (e. g. Costa & McCrae, 1985, 1987). The resulting test scores are also numbers. Therefore, psychological and educational measurement comprises the assignment of numbers to observational units according to some explicit rules[1]. This procedure is sometimes called scoring. There are many different approaches to score test takers on the basis of their response behaviour - sum scores, proportion correct scores, factor scores, etc. The use of a particular scoring method is typically justified by testing a corresponding measurement model. For example, if an IRT model - for example the Rasch model (Lord & Novick, 1968; Rasch, 1960) - is utilized and maximum likelihood estimates are used as test scores, the fit of the Rasch model to the observed data is tested. A good fit justifies the use of person parameter estimates of this model as test scores. No matter which model is chosen in a concrete application, the information used for scoring is given by persons' behaviour in response to a set of stimuli, which constitute the test. In most assessments stimuli are questions, statements, graphs, or tasks presented alongside with an instruction how to answer these items. The responses are scored. In the case of items in achievement tests, for example, the answer to an item can be correct, incorrect, or sometimes partially correct depending on the response format. The response pattern $y = y_1, \ldots, y_I$ consisting of the observed item scores y_i represents a person's response behaviour according to the test. In this work

[1] These numbers can also represent ordered ore unordered categories indicating different types of persons (e. g. latent class analysis; Rost, 2004) or skill levels (e. g. cognitive diagnostics models; von Davier, 2005; von Davier, DiBello, & Yamamoto, 2008).

particular stimuli of a test are not considered. For brevity, the term item i denotes the random variables Y_i (see below). Typically a test consists of more than a single item. Accordingly, if there are $I > 1$ items the response pattern $\boldsymbol{Y} = Y_1, \ldots, Y_I$ is an I-dimensional manifest variable. Random variables are defined with respect to particular probability space representing a concrete random experiment. In fact, most models of measurement theories such as CTT and IRT are probabilistic models. The term probabilistic refers at least to two aspects. First, the administration of a psychological or educational test is conceptualized as a random experiment (e. g. Steyer & Eid, 2001). Second, the test scores are considered to be fallible measures of latent unobserved variables constructed in measurement models. The relationships between the latent variables and manifest items or test scores are considered to be stochastic, which is formalized by the specification of linear or nonlinear regressions. In the subsequent Section the issue of latent variables in measurement models will be discussed in more detail.

Random experiment in psychological and educational testings Based on these considerations the random experiment that formally underlies psychological and educational assessments can be explicitly described considering the issue of potential missing data. The random experiment is:

(a) Draw randomly a person from the population under study.

(b) Observe the values of all the covariates $Z_1, \ldots, Z_J$.

(c) Administer the test consisting of I test stimuli. If item i is answered by the test taker observe the respective item score y_i and assign $D_i = 1$. If item i is missing assign $D_i = 0$.

This random experiment is formally represented by the probability space $(\Omega, \mathfrak{A}, P)$ (Steyer, 2002; Steyer & Eid, 2001; Steyer, Nagel, Partchev, & Mayer, in press). Compared to the random experiment described in the previous section (see Equation 2.1), additional random variables are involved in educational and psychological measurement. First, the person variable $U\colon \Omega \to \Omega_U$ is introduced since test takers are randomly selected. Second, a test consists usually of many items, each a random variable $Y_i\colon \Omega \to \Omega_{Y_i}$. Accordingly, the response pattern is the I-dimensional random variable $\boldsymbol{Y}\colon \Omega \to \Omega_{\boldsymbol{Y}}$. The response indicator variables are also random variables $D_i\colon \Omega \to \Omega_{D_i}$ on the same probability space with $\Omega_{D_i} = \{0, 1\}$ (see Equation 2.2). All response indicators taken together yield the missing indicator vector $\boldsymbol{D}\colon \Omega \to \Omega_{\boldsymbol{D}}$, which is also an I-dimensional random variable.

Finally, the J covariates Z_j: $\Omega \rightarrow \Omega_{Z_j}$ are combined to the multidimensional covariate $\boldsymbol{Z}$: $\Omega \rightarrow \Omega_{\boldsymbol{Z}}$. Based on this set of random variables the set of possible outcomes in a single unit trial is

$$\begin{aligned} \Omega &= \Omega_U \times \Omega_{Z_1} \times \ldots \times \Omega_{Z_J} \times \Omega_{D_1} \times \ldots \times \Omega_{D_I} \times \Omega_{Y_1} \times \ldots \times \Omega_{Y_I} \\ &= \Omega_U \times \Omega_{\boldsymbol{Z}} \times \Omega_{\boldsymbol{D}} \times \Omega_{\boldsymbol{Y}}. \end{aligned} \tag{2.7}$$

In most but not all educational and psychological measurements covariates are present. Therefore, sometimes a second slightly different random experiment will be considered in this work as well, which does not include covariates $Z_1, \ldots, Z_J$. This random experiment can be described as

(a) Draw randomly a person from the population under study.

(c) Administer the test consisting of I test stimuli. If item i is answered by the test taker observe the respective item score y_i and assign $D_i = 1$. If item i is missing assign $D_i = 0$.

The corresponding set of possible outcomes is

$$\begin{aligned} \Omega &= \Omega_U \times \Omega_{D_1} \times \ldots \times \Omega_{D_I} \times \Omega_{Y_1} \times \ldots \times \Omega_{Y_I} \\ &= \Omega_U \times \Omega_{\boldsymbol{D}} \times \Omega_{\boldsymbol{Y}}. \end{aligned} \tag{2.8}$$

In applications, parameters are aimed to be estimated based on realized data. A data set with N rows, which refers to the response pattern of the observational units, is the realization of a sample of size N. Hence, the single unit trail as described above (see Equation 2.7 and 2.8) needs to be repeated N times.

Taxonomy of missing data in the context of psychological and educational measurement In Section 2.1 the classification of the missing data mechanisms was introduced. All the definitions used here rest upon the conditional distributions $P(D \mid Y, \boldsymbol{Z})$. In educational and psychological measurement, however, not a single variable Y is considered but an I-dimensional random variable $\boldsymbol{Y}$ implying that the response indicator variable $\boldsymbol{D}$ is multivariate as well. It seems straightforward defining the missing data mechanisms on the basis of the conditional distributions of $\boldsymbol{D}$ given $(\boldsymbol{Y}, \boldsymbol{Z})$. However, in the case of multidimensional $\boldsymbol{Y}$, the case is more complex than this. Consider a very short test consisting of two items $\boldsymbol{Y} = (Y_1, Y_2)$. Additionally, there is a covariate $\boldsymbol{Z}$ that is stochastically

independent of Y_1 and Y_2. There is no missing data mechanism with respect to Y_1, so that $P(D_1 = 1) = 1$. The probability of missing Y_2 depends stochastically on Y_1, so that $P(D_2 = 1 \mid Y_1) \neq P(D_2 = 1)$ and $P(D_2 = 1 \mid \boldsymbol{Y}) = P(D_2 = 1 \mid Y_1)$. This implies $\boldsymbol{D} \not\perp (\boldsymbol{Y}, \boldsymbol{Z})$ and $\boldsymbol{D} \not\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. Hence, the occurrence of missing data with respect to $\boldsymbol{Y}$ is unconditionally and conditionally stochastically dependent on $\boldsymbol{Y}$ itself. One might be tempted to conclude that the missing data mechanism is NMAR. However, in this example follows $D_2 \perp Y_2 \mid (\boldsymbol{Z}, Y_1)$, since item Y_1 is never missing. $\boldsymbol{D}$ depends merely on completely observable variables. Hence, the missing data mechanism is MAR. This trivial example reveals that the definitions of missing data mechanisms are not trivial in the case of multidimensional variables. The approach chosen here is to define the nonresponse mechanisms with respect to each single item Y_i in a first step, and then to derive appropriate definitions for missingness with respect to $\boldsymbol{Y}$ in a second step.

Returning to the motivating example, the distinctive characteristics of the nonresponse mechanisms become apparent. Recall that the missingness of item Y_2 depends on a fully observed item Y_1 that would never suffer from missing data in application. Apart from branched testing, such items rarely exist in real psychological and educational testing. Rubin's definition does not require that any variables exist that need to be necessarily observed. The crucial point is that the probability of item nonresponses of a randomly chosen observational unit depends exclusively on observable variables. If so, the missing data mechanism is MAR. If the probability of missing responses is independent of any variable Y_i or Z_j considered in the random experiment the nonresponse mechanism is MCAR. Finally, if the probability of missingness depends on unobserved variables considered in the random experiment, the missing data mechanisms is NMAR. If all items Y_i have a probability to be missing in application, so that $P(D_i = 1) < 1$ (for all $i = 1, \ldots, I$), then it is also random which items will be observed and which will be missing. For that reason Rubin partitioned the complete data $\boldsymbol{y}$ into the observed part $\boldsymbol{y}_{\boldsymbol{obs}}$ and the missing part $\boldsymbol{y}_{\boldsymbol{mis}}$. Accordingly, here the item vector is partitioned into $\boldsymbol{Y} = (\boldsymbol{Y}_{\boldsymbol{obs}}, \boldsymbol{Y}_{\boldsymbol{mis}})$. In order to define the missing data mechanisms with respect to a single item Y_i, the (un)conditional stochastic relationships between item i, the covariate $\boldsymbol{Z}$, and observable and unobservable items $Y_{k \neq i}$ are considered. Let $\boldsymbol{Y}^{-\boldsymbol{i}} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_I)$ be the item vector without item i that can also be partitioned into $\boldsymbol{Y}^{-\boldsymbol{i}} = \left(\boldsymbol{Y}^{-\boldsymbol{i}}_{\boldsymbol{obs}}, \boldsymbol{Y}^{-\boldsymbol{i}}_{\boldsymbol{mis}}\right)$. This partition can be different for every respondent. In fact there are as many partitions of $\boldsymbol{Y}^{-\boldsymbol{i}}$ as possible missing patterns. For the case of I items 2^{I-1} missing pattern with respect to $\boldsymbol{Y}^{-\boldsymbol{i}}$ exist. The nonresponse mechanisms with respect to single items Y_i can be defined as:

- The missing data mechanism w.r.t. Y_i is *missing completely at random (MCAR)* if

$$P(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}) \quad = \quad P(D_i = 1). \tag{2.9}$$

- The missing data mechanism w.r.t. Y_i is *missing at random (MAR) given* $(\boldsymbol{Y}, \boldsymbol{Z})$, if

$$P(D_i = 1 \mid Y_i) \neq P(D_i = 1), \tag{2.10}$$

  and

$$P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}\right). \tag{2.11}$$

- Two special cases can be considered in the context of psychological and educational measurement satisfying Equation 2.11. The missing data mechanism w.r.t. Y_i is called *missing at random given* $\boldsymbol{Z}$, if Equation 2.10 holds and

$$P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Z}\right). \tag{2.12}$$

- In contrast, the missing data mechanism w.r.t. Y_i is called *missing at random given* $\boldsymbol{Y}$, if Equation 2.10 holds and

$$P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}\right). \tag{2.13}$$

- The missing data mechanism w.r.t. Y_i is called *not missing at random (NMAR)* or non-ignorable if

$$P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) \neq P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}\right). \tag{2.14}$$

Based on these definitions the nonresponse mechanism of the complete item vector $\boldsymbol{Y}$ can be defined. It is important to note that the missing data mechanisms can vary across the items within a single test. That is, nonresponses of some items can be MCAR or MAR while missing responses to other items can be nonignorable. The following definitions of the missing data mechanisms with respect to the complete measurement instrument $\boldsymbol{Y}$ account for potentially different coexisting nonresponse mechanisms with regard to single item. The definitions are built on the definitions of the missing data mechanism with respect to Y_i. The following missing data mechanisms with respect to $\boldsymbol{Y}$ can be

defined.

- The missing data mechanism with respect to $\boldsymbol{Y}$ is *missing completely at random (MCAR)* if Equation 2.9 holds for all items Y_i, implying that

$$\boldsymbol{D} \perp (\boldsymbol{Y}, \boldsymbol{Z}). \tag{2.15}$$

- The missing data mechanism with respect to $\boldsymbol{Y}$ is *missing at random (MAR) given* $(\boldsymbol{Y}, \boldsymbol{Z})$ if at least one of the following two conditions hold true:

  1. $\exists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}\right)\right)$
  2. $\exists\, (i, j)$ with $(i \neq j)$
     $\left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}\right)\right)$, and
     $\left(P(D_j = 1 \mid Y_i) \neq P(D_j = 1) \wedge P\left(D_j = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_j = 1 \mid \boldsymbol{Z}\right)\right)$.

  Additionally, it is required that
  $\nexists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) \neq P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}\right)\right)$.
  Hence, the probability of missingness of items depends on observable items and the fully observable covariate $\boldsymbol{Z}$ but not on unobserved items. Hence,

$$\boldsymbol{D} \not\perp (\boldsymbol{Y}, \boldsymbol{Z}), \text{ and} \tag{2.16}$$

$$\boldsymbol{D} \perp \boldsymbol{Y}_{\boldsymbol{mis}} \mid (\boldsymbol{Y}_{\boldsymbol{obs}}, \boldsymbol{Z}). \tag{2.17}$$

- The missing data mechanism with respect to $\boldsymbol{Y}$ is *missing at random (MAR) given* $\boldsymbol{Z}$, if two conditions hold:

  1. $\exists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Z}\right)\right)$
  2. Additionally, it is required that
     $\nexists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) \neq P\left(D_i = 1 \mid \boldsymbol{Z}\right)\right)$.

  In this case, Equations 2.16 and 2.17 hold true. Hence, this nonresponse mechanism is a special case of the MAR mechanism given $(\boldsymbol{Y}, \boldsymbol{Z})$. Additionally, it applies conditional stochastic independence

$$\boldsymbol{D} \perp \boldsymbol{Y} \mid \boldsymbol{Z}. \tag{2.18}$$

- The missing data mechanism with respect to $\boldsymbol{Y}$ is *missing at random given (MAR)* $\boldsymbol{Y}$, if two conditions hold:

1. $\exists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) = P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}\right)\right)$
2. Additionally, it is required that
   $\nexists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) \neq P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}\right)\right)$.

Again, Equations 2.16 and 2.17 hold true. In fact, this nonresponse mechanism is the second spacial case of the MAR condition given $(\boldsymbol{Y}, \boldsymbol{Z})$. Since the probability of item nonresponses depends only on observable items it applies

$$\boldsymbol{D} \perp \boldsymbol{Y}_{\boldsymbol{mis}} \mid \boldsymbol{Y}_{\boldsymbol{obs}}. \tag{2.19}$$

- Finally, the missing data mechanism w.r.t. $\boldsymbol{Y}$ is *non-ignorable* or *not missing at random (MNAR)* if

  $$\exists\, i \left(P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \wedge P\left(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}\right) \neq P\left(D_i = 1 \mid \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}\right)\right),$$

  implying that

  $$\boldsymbol{D} \not\perp \boldsymbol{Y}_{\boldsymbol{mis}} \mid (\boldsymbol{Y}_{\boldsymbol{obs}}, \boldsymbol{Z}). \tag{2.20}$$

In the remainder of this work the terms MCAR, MAR, and NMAR will refer to these definitions of the missing data mechanisms that are defined either with respect to single items Y_i or with respect to the item vector $\boldsymbol{Y}$. In cases where one or more of the three MAR conditions - MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$, $\boldsymbol{Z}$ or $\boldsymbol{Y}$ - are discussed, the nonresponse mechanism will sometimes simply be called MAR to facilitate the reading. Depending on the specific context, one of the three MAR conditions will be addressed, otherwise statements will apply to all three MAR conditions. It is important to note that two MAR conditions cannot be defined with respect to the random experiment described by Equation 2.8. If no covariate $\boldsymbol{Z}$ is involved only three missing data mechanism exist with respect to Y_i and the item vector $\boldsymbol{Y}$. These are (a) MCAR, (b) MCAR given $\boldsymbol{Y}$, and (c) NMAR. The definitions of these three nonresponse mechanisms as introduced above hold in this case as well. Simply the covariate $\boldsymbol{Z}$ needs to be omitted from the equations. Note that these definitions do not require that all items have the same nonresponse mechanism. If the missing data mechanism with respect to $\boldsymbol{Y}$ is called nonignorable, that implies that there is *at least* one item having a missing data mechanism that is NMAR. The remaining items can have nonresponse mechanism that are MAR or MCAR. Similarly, given the missing data mechanism with respect to $\boldsymbol{Y}$ is MAR, some of the items can have a nonresponse

mechanism that is even MCAR.

Relation to Rubin's definitions As previously stated Rubin's approach to handle missing data is deeply inspired by Bayesian thinking. This is also reflected in his definitions of nonresponse mechanisms. Here the relation between the previously introduced definitions used throughout this work and Rubin's definitions (1976) will be briefly explained. Here in this work the nonresponse mechanisms have been defined with respect to the single unit trial. Rubin, however, considers a $N \times I$ data matrix $\boldsymbol{Y} = \boldsymbol{y}$, sometimes called the complete data matrix. He partitions the data matrix into an observed part $\boldsymbol{y}_{obs}$ and an unobserved or missing part $\boldsymbol{y}_{mis}$. Hence $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$. Table 2.1 shows an example with the sample size $N = 5$ and $I = 3$ dichotomous items Y_1, Y_2, and Y_3. In the general case ii is not distinguished between dependent and independent variables or covariates. Hence, Y_3 in the data example (see Table 2.1) is analogous to the fully observed covariate $\boldsymbol{Z}$ used in the definitions here. Rubin defined the missing data mechanisms based on partitions $\boldsymbol{y}_{obs}$ and $\boldsymbol{y}_{mis}$ of

Table 2.1: Example of the Partitions of Complete Data $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$ and the Corresponding Response Indicator Matrix $\boldsymbol{d}$.

				observed						unobserved		
n	$\boldsymbol{y}$			$\boldsymbol{d}$			$\boldsymbol{y}_{obs}$			$\boldsymbol{y}_{mis}$		
1	1	0	1	1	1	1	1	0	1	*	*	*
2	0	1	1	1	0	1	0	*	1	*	1	*
3	1	0	0	1	0	1	1	*	0	*	0	*
4	0	1	0	0	0	1	*	*	0	0	1	*
5	0	1	1	0	1	1	*	1	1	0	*	*

* indicates nonexistent elements in the respective partition.

the complete data matrix and the missing pattern $\boldsymbol{D} = \boldsymbol{d}$ using factorization methods that are close to factorization of joint probability or density functions and likelihood functions (e. g. Barndorff-Nielsen, 1976; Cox & Wermuth, 1999; Cramér, 1949). Unfortunately, in most missing data literature it is not clearly distinguished between data and random variables. Often it is written that missingness depends on observed or unobserved data. In turn unobserved data are often said to be conditionally independent of missingness given the observed data, or the distribution of missing data is discussed. This may be partly due to the fact that the terms distribution and posterior distribution are often used synonymously. However, from a frequentist?s perspective, the terms probability and density function are meaningful with respect to random variables but not to data. However, this work is not intended to criticize or to correct commonly used terminology. This section only aims to connect the previously introduced definitions to Rubin's framework. Rubin assumed a

joint probability or density function $g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d})$ of the response indicator matrix and the complete data matrix that can be written as $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d})$. The joint probability can be factored into

$$\begin{aligned} g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}) \\ &= g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}). \end{aligned} \tag{2.21}$$

Typically, two parameter vectors $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$ are introduced. $\boldsymbol{\iota}$ is the vector of parameters referring to the target model of substantial interest. $\boldsymbol{\phi}$ is a parameter vector indexing the missing data model. For example, if $\boldsymbol{D}$ could be appropriately modelled by an I-variate logistic regression $P(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y})$, $\boldsymbol{\phi}$ would be the vector of logistic regression coefficients and intercepts. Including these parameterizations, Equation 2.21 can be written as

$$\begin{aligned} g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \qquad (2.22) \\ &= g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota}). \qquad (2.23) \end{aligned}$$

Given the missing data mechanism is MCAR the probability of a missing pattern equal to that observed in the sample is independent of any random variable in the model. Hence, Rubin defined the missing data mechanism to be MCAR if

$$\begin{aligned} g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \\ &= g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota}). \end{aligned} \tag{2.24}$$

In this work the missing data mechanisms were defined with regard to the single unit trial. It can be shown that the resulting data matrices resulting from N repetition of the single unit trial will have the properties described by Rubin if $N \to \infty$. If there are N independent single unit trials, $\boldsymbol{Y}$ and $\boldsymbol{D}$ become $(N \times I)$-dimensional random variables with a joint distribution $g(\boldsymbol{Y}, \boldsymbol{D})$. Under an explicit joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$ indexed by $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$ the joint probability is

$$\begin{aligned} g(\boldsymbol{Y}, \boldsymbol{D}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= \prod_{n=1}^{N} g(\boldsymbol{Y}_n, \boldsymbol{D}_n; \boldsymbol{\iota}, \boldsymbol{\phi}) \\ &= \prod_{n=1}^{N} g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}, \boldsymbol{D}_n; \boldsymbol{\iota}, \boldsymbol{\phi}). \end{aligned} \tag{2.25}$$

If the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MCAR as defined previously (see Equation 2.15) it follows

$$\begin{aligned} g(\boldsymbol{Y}, \boldsymbol{D}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= \prod_{n=1}^{N} g(\boldsymbol{D}_n; \boldsymbol{\phi}) g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\iota}). \\ &= \underbrace{\prod_{n=1}^{N} g(\boldsymbol{D}_n; \boldsymbol{\phi})}_{g(\boldsymbol{D};\boldsymbol{\phi})} \underbrace{\prod_{n=1}^{N} g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\iota})}_{g(\boldsymbol{Y}_{obs},\boldsymbol{Y}_{mis},\boldsymbol{\iota})}. \end{aligned} \tag{2.26}$$

Following Rubin the probability of each event $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{D} = \boldsymbol{d}$ can be written as shown in Equation 2.24.

Similarly, Rubin defined the nonresponse mechanism to be MAR if

$$\begin{aligned} g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \\ &= g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota}). \end{aligned} \tag{2.27}$$

In this work, three different MAR conditions are distinguished due to the distinction between manifest variables in the measurement model and covariates. However, one can directly compare the definition of the nonresponse mechanism with respect to $\boldsymbol{Y}$ based on the single unit trial without covariates (see Equation 2.8). In this case, the missing data mechanism with respect to $\boldsymbol{Y}$ was defined to be MAR if $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ holds in the single unit trial. Again, in a sample of N single unit trials $\boldsymbol{Y}$ and $\boldsymbol{D}$ become $(N \times I)$ dimensional random matrices with a joint distribution. Analogous to Equation 2.25 the joint distribution under a model $(\boldsymbol{\iota}, \boldsymbol{\phi})$ is

$$\begin{aligned} g(\boldsymbol{Y}, \boldsymbol{D}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= \prod_{n=1}^{N} g(\boldsymbol{D}_n \,|\, \boldsymbol{Y}_{n;obs}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\iota}). \\ &= \underbrace{\prod_{n=1}^{N} g(\boldsymbol{D}_n \,|\, \boldsymbol{Y}_{n;obs}; \boldsymbol{\phi})}_{g(\boldsymbol{D}\,|\,\boldsymbol{Y}_{obs};\boldsymbol{\phi})} \underbrace{\prod_{n=1}^{N} g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\iota})}_{g(\boldsymbol{Y}_{obs},\boldsymbol{Y}_{mis},\boldsymbol{\iota})}. \end{aligned} \tag{2.28}$$

Hence, the probability of each event $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{D} = \boldsymbol{d}$ under the defined missing data mechanisms can be written according to Equation 2.27.

Finally, the missing data mechanism with respect to $\boldsymbol{Y}$ was defined to be nonignorable based on the single unit trial if $\boldsymbol{D} \not\perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ (cf. Equation 2.20). The joint probability of

$(N \times I)$ dimensional random matrices $\boldsymbol{Y}$ and $\boldsymbol{D}$ is then

$$
\begin{aligned}
g(\boldsymbol{Y}, \boldsymbol{D}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= \prod_{n=1}^{N} g(\boldsymbol{D}_n \mid \boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\iota}). \qquad (2.29) \\
&= \underbrace{\prod_{n=1}^{N} g(\boldsymbol{D}_n \mid \boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\phi})}_{g(\boldsymbol{D} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi})} \underbrace{\prod_{n=1}^{N} g(\boldsymbol{Y}_{n;obs}, \boldsymbol{Y}_{n;mis}; \boldsymbol{\iota})}_{g(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\iota})}.
\end{aligned}
$$

Therefore, the probability of each event $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{D} = \boldsymbol{d}$ as given by Equation 2.23 cannot be further simplified. If the missing data mechanism is MCAR or MAR, missingness depends only on observable variables. However, if the nonresponse mechanism is NMAR the probability of the occurence of missing data also depends on unobserved variables. The term ignorable and nonignorable missing data are also commonly used. These statements are strictly speaking only meaningful with respect to a joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$ indexed by $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$. The nonresponse mechanism is said to be *ignorable* if the nonresponse mechanism is MAR or MCAR, and the joint parameter space $\Omega_{(\boldsymbol{\iota},\boldsymbol{\phi})}$ of $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$ can be written as a cartesian product $\Omega_{\boldsymbol{\iota}} \times \Omega_{\boldsymbol{\phi}}$ (Little & Rubin, 2002; Rubin, 1976). If the missing data mechanism is NMAR and/or the parameter space is restricted so that $\Omega_{(\boldsymbol{\iota},\boldsymbol{\phi})} \neq \Omega_{\boldsymbol{\iota}} \times \Omega_{\boldsymbol{\phi}}$, the nonresponse mechanism is called *nonignorable*.

Informative and noninformative missingness Alternatively, the terms *informative* and *noninformative* drop out or missingness are commonly used to indicate ignorable and nonrignorable missing data mechanisms, respectively. In fact these terms may better reflect the problem of missing data. If the nonresponse mechanism is ignorable, the missing data are noninformative. Hence, missingness itself does not provide additional information with respect to estimable parameters of interest. In other words, over and above the observed data $\boldsymbol{y}_{obs}$, the missing pattern $\boldsymbol{d}$ is not informativ with respect to $\boldsymbol{\iota}$. In contrast, if the missing data mechanism is nonignorable, missing data are informative. In this case $\boldsymbol{d}$ contains information about unknown parameters in $\boldsymbol{\iota}$. Simply speaking the observed data are non-representative. Inference exclusively resting upon $\boldsymbol{y}_{obs}$ is then potentially biased, since the sampling distribution of estimates $\hat{\boldsymbol{\iota}}$ of $\boldsymbol{\iota}$ differs even if $N \to \infty$.

Multiple imputation (MI) is particularly suited to illustrate the difference between ignorable (noniformative) and nonrignorable (informative) missingness. Multiple imputation has become a widely used method to handle missing data occurring under an ignorable nonresponse mechanism (Enders, 2010; Graham, 2009; Little, 1988a; Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002). MI is also valuable for a better understand-

ing of the missing data terminology used in Rubin's framework, since it relies heavily on Bayesian statistics. MI is a multi-step data augmentation method. In the imputation phase missing data $\boldsymbol{y}_{mis}$ are replaced by plausible values (PV) estimated from the observed data $\boldsymbol{y}_{obs}$. In order to account sufficiently for variance due to imputations, multiple, typically $m = 5$, data sets are generated. In the analysis phase, each of the m filled-in data sets can be analyzed with statistical standard methods for complete data. In the final pooling phase the results of the m analyses are combined to yield single point estimates and associated statistics (e. g. standard errors). The joint distribution $g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ can be factored in various ways. Alternative to Equation 2.21 the joint distribution can be factorized as follows:

$$\begin{aligned} g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}_{mis}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}) \\ &= g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}_{mis}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) \\ &\quad \cdot g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_{obs}) \end{aligned} \tag{2.30}$$

The first factor $g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}_{mis})$ refers to the *predictive distribution* of the missing data conditional on all observed data $\boldsymbol{y}_{obs}$ and $\boldsymbol{d}$ (Little & Rubin, 2002; Rubin, 1987; Schafer, 1997). Under MI, random draws from the predictive distribution are drawn to fill in the incomplete observed data. The vector $\boldsymbol{\iota}_{mis}$ consists of the parameters of the imputation model, which could be regression coefficients, residual variances, and covariances. $\boldsymbol{\iota}_{obs}$ is a parameter vector that describes the distribution of the observables, whereas $\boldsymbol{\phi}$ refers to the missing data model conditional on observed values $\boldsymbol{y}_{obs}$. Recall that in Bayesian statistics there is no difference between unknown model parameters aimed to be estimated and missing data $\boldsymbol{y}_{mis}$ (Gelman, Carlin, Stern, & Rubin, 2003). These unobserved quantities have a distribution due to uncertainty about their true values. In this work the missing data mechanism was defined by conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ (see Equation 2.18). This condition also holds true if the missing data mechanism is MCAR. Hence, if the missing data mechanism is ignorable Equation 2.30 can be simplified and rearranged yielding

$$\begin{aligned} g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_{mis}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_{obs}) \\ &\quad \cdot g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) \\ &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) \end{aligned} \tag{2.31}$$

Two results are important to note: (a) If the nonresponse mechanism is ignorable then missingness stochastically depends only on observable variables $\boldsymbol{Y}_{obs}$, and (b) the same applies to the the predictive distribution of unobservable variables $\boldsymbol{Y}_{mis}$, which also depends on $\boldsymbol{Y}_{obs}$ exclusively. Hence, missingness indicated by $\boldsymbol{D} = \boldsymbol{d}$ can be *ignored*, since it is *not informative* with respect to unobserved variables over and above observable variables $\boldsymbol{Y}_{obs}$. A correctly specified imputation model-based on $\boldsymbol{Y}_{obs}$ ensures unbiased and consistent parameter estimates and valid sample based inference. In contrast, if the missing data mechanism is NMAR, different missing patterns $\boldsymbol{D} = \boldsymbol{d}$ are associated with different distributions of $\boldsymbol{Y}_{mis}$ even conditional on observable variables $\boldsymbol{Y}_{obs}$. Hence, $\boldsymbol{D}$ cannot be ignored in the imputation model since it is *informative* with respect to the predictive distribution of $\boldsymbol{Y}_{mis}$. Unfortunately, the parameters $\boldsymbol{\iota}_{mis}$ of the imputation model are difficult to estimate without strong assumptions in the case of nonignorable missing data. That is why the application of MI is typically limited to ignorable missing data.

Starting from these considerations, the implications of the different missing data mechanisms with respect to latent variables indicated by $\boldsymbol{Y}$ will be examined analytically in the following section. The idea is quite simple. If the distribution of $\boldsymbol{Y}$ depends on $\boldsymbol{D}$ the distribution of the latent variables indicated by $\boldsymbol{Y}$ potentially depends on $\boldsymbol{D}$ as well.

2.3 Implications With Respect to Underlying Variables

Before the implications of the different missing data mechanisms with respect to latent variables will be discussed, some general notes are made with respect to commonly used terms in measurement theory such as constructs, latent variables and scores. A comprehensive introduction to measurement theory is far beyond the scope of this work but can be found in volumes such as Lord & Novick (1968), Rost (2004), Steyer (1989, 2001) and Steyer & Eid (2001). This introduction merely aims to clarify how relevant terms are used here in this work.

The term test scores, or simply scores, has already been clarified in section 2.2. They result from scoring or scaling procedures. Independent of the particular used scoring method, successful measurement means that the relations between scores reflect relations between observational units regarding the characteristics of interest. In most psychological and educational assessments scaling procedures are model-based. That is, a measurement model is proposed that models explicate the relation between latent and observable variables (Steyer & Eid, 2001). The validity of measurement models has typically testable implications. The model fit assessed by many different indices and test

statistics can be used to justify a particular scoring method empirically. Most probabilistic measurement models used in educational and psychological assessments are models of Classical Test Theory (Borsboom & Mellenbergh, 2002) (Skrondal & Rabe-Hesketh, 2004; Steyer & Eid, 2001; Steyer, 2002) or Item Response Theory (de Ayala, 2009; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Skrondal & Rabe-Hesketh, 2004; Steyer & Eid, 2001). Within each of both theories a considerable number of models have been developed. For example, in CTT the parallel test model, the model of τ-equivalent variables, and the model of τ-congeneric variables are well known (e. g. Steyer & Eid, 2001). τ refers to the true score variables or simply true scores defined as the expected scores given the person. A more formal definition will be given below. There is a growing number of IRT models that can be classified in different ways (D. M. Thissen & Steinberg, 1986). It is far beyond the scope of this work to provide a comprehensive overview of these models. However, uni- and multidimensional one-, two, and three-parametric logistic or probit models are the most frequently used IRT models for dichotomous items. Despite notable differences between the CTT and IRT, both have a lot in common. Measurement models of both theoretical approaches are stochastic models that define manifest variables to be dependent variables in multivariate regressions on latent variables that are not directly observable. In CTT these regressions are linear whereas non-linear regressions are used in IRT.

Although latent variable models are common in psychology and educational research and the term latent variable is routinely used, there are considerably different views about what a latent variable is (e. g. Bollen, 2002; Borsboom, Mellenbergh, & Van Heerden, 2003; Borsboom, 2008; J. R. Edwards & Bagozzi, 2000). There is a common belief that latent variables exist without a model and only need to be measured. As Steyer emphasized (Steyer & Eid, 2001), latent variables in measurement models are not a priori existent but are constructed in this particular measurement model. They are mathematically well-defined with respect to a particular considered random experiment. It is important to distinguish between constructs of theoretical interest and latent variables. As Borsboom and Mellenbergh (2002) complained, many authors simply equate latent variables to the construct of interest. However, this is not only incorrect but inconsistent with some definitions of CTT and IRT (Lord & Novick, 1968). For that reason Borsboom and Mellenbergh distinguish between *latent variable scores* and *construct scores* to highlight that the values of a latent variable in a specific measurement model are not per se equal to the value on the construct of interest. This is intuitive considering the fact that, theoretically, an infinite number of models exist to explain relations between variables. In each of these

models the individual values of latent variables might differ whereas the construct scores do not. Each model might assign different scores to the same observational unit due to its specific response behaviour. Furthermore, latent variables in measurement models of CTT and IRT are syntactic concepts (Borsboom & Mellenbergh, 2002). That is, they are well-defined in mathematical terms. In contrast, construct scores refer to a semantic concept which originates from substantial theories. The relation between construct scores and latent variables scores is an issue of validity. As Borsboom and Mellenbergh stated, the definition of constructs or construct scores is „generally difficult“. Consequently, the concept of validity and the process of test validation is controversially discussed (Borsboom, Mellenbergh, & Van Heerden, 2004; Borsboom, 2006). Nevertheless, in this work it is unavoidable to address the issue of test validity. It will be repeatedly demonstrated that item nonresponses and their treatment can affect the construction of latent variables in measurement models such that the substantial interpretation of the latent variable can change. This fact justifies the statement that validity is threatened by missing data and their treatment even if a completely satisfying definition of validity is still lacking. Construct scores will not be further considered here since, to the best knowledge of the author, they cannot sufficiently be captured by mathematical terms yet. In order to study and to illustrate effects of missing data and their handling, fictional and simulated data examples will be used. The terms latent ability, latent proficiency, or latent trait refers to the latent variable $\boldsymbol{\xi}$ constructed in a measurement model and not to constructs or construct scores.

In the remainder of this section the implications of the different nonresponse mechanisms with respect to the distribution of latent variables underlying observed and missing data will be studied. In the previous section the typology of missing data was introduced. The nonresponse mechanisms were exclusively defined by the conditional and unconditional stochastic dependencies between manifest variables Y_i, D_i, $\boldsymbol{Y}$, $\boldsymbol{D}$ and the covariate $\boldsymbol{Z}$. Latent variables underlying $\boldsymbol{Y}$ were not included in the definitions of missing data mechanisms. The reason is that latent variables are not a priori existent but constructed based on $\boldsymbol{Y}$ in a concrete random experiment. Hence, latent variables are always missing. Accordingly, distributional parameters and individual values of latent variables needs to be inferred from the data $\boldsymbol{y}$. In IRT the values of latent variables are person parameters aimed to be estimated with respect to each test taker. Hence, the values of latent variables or at least distributional parameters of latent variables are part of the parameter vector $\boldsymbol{\iota}$ and are, therefore, not used her for defining nonresponse mechanisms. Instead, the implications of the different missing data mechanisms with respect to distributions of latent variables are examined. The aim is to demonstrate analytically why item and person pa-

rameter estimates are potentially biased when the missing data are NMAR. At first, the true scores τ_i are considered followed by the latent variable denoted by $\boldsymbol{\xi}$.

Implications with respect to true score variables

In CTT the manifest variables Y_i can be decomposed in a true score variable τ_i and the residual ε_i. Although this decomposition is distinctive of CTT it does not contradict with IRT. Regardless of whether the Y_i is continuous or discrete it can be written as

$$\begin{aligned} Y_i &= \tau_i + \varepsilon_i && (2.32) \\ &= E(Y_i \mid U) + \varepsilon_i. && (2.33) \end{aligned}$$

The true score τ_i is defined as the regression $E(Y_i \mid U)$ of the manifest variable on the unit variable U. ε_i is the residual of the regression $E(Y_i \mid U)$, defined as the difference $Y_i - \tau_i$. The true score variable τ_i is a function $f_i(U)$ of the unit variable U. The conditional expected values $E(Y_i \mid U = u) = \tau_i(u)$ are the individual expected scores of a person u with respect to test/item Y_i. In the case of single items Y_i the true score will also be called the expected item score. τ_i is well defined even if the item or the test was never presented. Hence, the conditional distribution of $g(Y_i \mid D_i = 0)$ with the expected value $E(Y_i \mid D_i = 0)$ can be considered without theoretical inconsistencies. Furthermore, the regression $E(Y_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z})$ of test/items i on the response indicator and other test/items and the covariate can be considered. Inserting the right-hand side of Equation 2.32 into this regression yields

$$\begin{aligned} E(Y_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z}) &= E(\tau_i + \varepsilon_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z}) && (2.34) \\ &= E(\tau_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z}) + E(\varepsilon_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z}). && (2.35) \end{aligned}$$

When assuming regressive independence for the residual term, i.e, $\varepsilon_i \vdash (D_i, \boldsymbol{Y^{(-i)}}, \boldsymbol{Z})$ it follows that

$$E(Y_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z}) = E(\tau_i \mid D_i, \boldsymbol{Y^{-i}}, \boldsymbol{Z}). \quad (2.36)$$

Thus, the expected score on item Y_i given the response indicator variable and all other variables is equal to the expectation of the true score variable τ_i given these variables.

Implications of MCAR with respect to true scores Using the definition of MCAR with respect to Y_i, due to symmetry properties, stochastic independence $P(D_i = 1 \mid \boldsymbol{Y}, \boldsymbol{Z}) =$

$P(D_i)$ (see Equation 2.9) implies:

$$(a) \quad P(D_i = 1 \mid Y_i) = P(D_i = 1) \tag{2.37}$$

$$(b) \quad g(Y_i \mid D_i = d_i) = g(Y_i) \tag{2.38}$$

$$(c) \quad \forall\, (\boldsymbol{y}^{-i}, z) \in \Omega_{\boldsymbol{Z}} \times \Omega_{\boldsymbol{Y}^{-i}} \; g(Y_i \mid D_i = 1, \boldsymbol{Y}^{-i} = \boldsymbol{y}^{-i}, \boldsymbol{Z} = z) = g(Y_i \mid \boldsymbol{Y}^{-i} = \boldsymbol{y}^{-i}, \boldsymbol{Z} = z) \tag{2.39}$$

Note that the definition of MCAR has no implications with respect to stochastic dependencies between items Y_i, the remaining items $\boldsymbol{Y}^{-i}$ and the covariate $\boldsymbol{Z}$. Of course, the distribution of math items Y_i, for example, can depend on background variables such as socioeconomic status or other items in the test. Such relationships remain unaffected by any missing data mechanism. However, under MCAR missingness of Y_i is stochastically independent of all considered variables. Therefore, responses to item Y_i depend not on the probability to respond to this item. Stochastic independence implies regressive independence. In combinations with the assumptions of regressive independence of the residual ε_i MCAR implies

$$E(\tau_i \mid D_i, \boldsymbol{Y}^{-i}, \boldsymbol{Z}) = E(\tau_i \mid \boldsymbol{Z}, \boldsymbol{Y}^{-i}) \tag{2.40}$$

$$E(\tau_i \mid D_i) = E(\tau_i). \tag{2.41}$$

Equations 2.37 and 2.41 show that Y_i is identically distributed regardless of the observational status. In application that means that the observed values of the item vector $y_{1i}, \ldots, y_{Ni}$ are a representative with respect to the unconditional distribution of Y_i. Similarly, conditional on $(\boldsymbol{Y}^{-i}, \boldsymbol{Z})$ the observations $y_{1i}, \ldots, y_{Ni}$ are representative with respect to the conditional distribution $g(Y_i \mid \boldsymbol{Y}^{-i} = \boldsymbol{y}^{-i}, \boldsymbol{Z} = z)$. Hence, the observed data $\boldsymbol{y_{i;obs}}$ on item i available for analyses are reduced by item non-responses but representative in all subgroups and the complete sample. Equation 2.41 shows that the performance on item Y_i is on average equivalent for test takers who answer the items and those who have missing responses. There is no systematic dropout with respect to the true scores. In the case of dichotomous items this means that persons with higher or lower probabilities of a correct answer are not more or less likely to not respond to item Y_i.

Implications of MAR w. r. t True Scores If the missing data mechanism w.r.t. to item Y_i is MAR, the probability of item nonresponse to item i is stochastically dependent on observed variables $\boldsymbol{Z}$ and $\boldsymbol{Y_{obs}}$ (see Equations 2.11 - 2.13). A defining characteristic that

distinguishes MAR w.r.t. Y_i from MCAR is unconditional stochastic dependence between the response indicator D_i and Y_i. Hence

$$P(D_i = 1 \mid Y_i) \neq P(D_i = 1) \quad \Rightarrow \quad g(Y_i \mid D_i = 1) \neq g(Y_i). \tag{2.42}$$

In contrast to the case of MCAR, the distribution of the manifest test variable Y_i is different depending on the status of missingness. Hence, in application, all realizations of Y_i originate from the distribution $g(Y_i \mid D_i = 1)$ that is different from the unconditional distribution $g(Y_i)$. The implications of Equation 2.42 with respect to the true score variables depends on the distribution of Y_i and the residual ε_i. Let Y_i be a continuous variable with ε_i normally distributed $N[0, Var(\varepsilon_i)]$. If independence $\varepsilon_i \perp D_i$ hold, then

$$(a) \quad g(Y_i \mid D_i = 1) \neq g(Y_i) \Rightarrow g(\tau_i \mid D_i = 1) \neq g(\tau_i) \tag{2.43}$$

$$(b) \quad Var(Y_i \mid D_i = 1) \neq Var(Y_i) \Rightarrow Var(\tau_i \mid D_i = 1) \neq Var(\tau_i) \tag{2.44}$$

$$(c) \quad E(Y_i \mid D_i = 1) \neq E(Y_i) \Rightarrow E(\tau_i \mid D_i = 1) \neq E(\tau_i) \tag{2.45}$$

This work focuses primarily on dichotomous variables Y_i which follow a Bernoulli distribution, with $E(Y_i) = P(Y_i = 1)$ and $Var(Y_i) = P(Y_i = 1)[1 - P(Y_i = 1)]$. Hence, the distribution of the test variable is sufficiently described by the probability $P(Y_i = 1)$. The expected value of Y_i is the unconditional probability of solving item Y_i that can be written as

$$\begin{aligned} P(Y_i = 1) &= E[P(Y_i = 1 \mid U)] \\ &= E(\tau_i). \end{aligned} \tag{2.46}$$

Additionally, for dichotomous Y_i the residual variance is

$$\begin{aligned} Var(\varepsilon_i) &= E[Var(\varepsilon_i \mid U)] \\ &= E[Var(Y \mid U)] \\ &= E[\tau_i(1 - \tau_i)]. \end{aligned} \tag{2.47}$$

Thus, the distribution of dichotomous items Y_i depends exclusively on the distribution of the true score variables. Consequently, given the missing data mechanism w.r.t. Y_i is MAR, Equation 2.10 implies $\varepsilon_i \not\perp D_i$ and that Equations 2.43 - 2.45 hold true without further assumptions.

Three different MAR definitions have been introduced here. Generally, MAR with

respect to Y_i has been defined as a conditional stochastic independence of response indicators D_i of $\left(Y_i, \boldsymbol{Y}_{\boldsymbol{mis}}^{-\boldsymbol{i}}\right)$ given observable variables $\left(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}\right)$. Equations 2.10 and 2.11 implies $\forall\ (\boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, z) \in \Omega_{\boldsymbol{Y}_{\boldsymbol{obs}}^{-i}} \times \Omega_{\boldsymbol{Z}}$

$$g(Y_i \,|\, D_i = 1, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z) = g(Y_i \,|\, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z) \tag{2.48}$$

$$E(\tau_i \,|\, D_i, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}) = E\left(\tau_i \,|\, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}\right) \tag{2.49}$$

There are two ignorable nonresponse mechanisms with respect to Y_i that are special cases of the MAR condition defined by Equations 2.10 and 2.11. If the missing data mechanisms with respect to Y_i is MAR given $\boldsymbol{Z}$, then Equation 2.12 implies $\forall\ z \in \Omega_{\boldsymbol{Z}}$

$$g(Y_i \,|\, D_i = 1, \boldsymbol{Z} = z) = g(Y_i \,|\, \boldsymbol{Z} = z) \tag{2.50}$$

$$E(\tau_i \,|\, D_i, \boldsymbol{Z}) = E(\tau_i \,|\, \boldsymbol{Z}). \tag{2.51}$$

Equivalently, if the nonresponse mechanism with respect to Y_i is MAR given $\boldsymbol{Y}$ as defined above, then Equation 2.13 implies $\forall\ \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} \in \Omega_{\boldsymbol{Y}_{\boldsymbol{obs}}^{-i}}$

$$g(Y_i \,|\, D_i = 1, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}) = g(Y_i \,|\, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}) \tag{2.52}$$

$$E(\tau_i \,|\, D_i, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}) = E(\tau_i \,|\, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}) \tag{2.53}$$

Thus, the distinctive feature of MAR is that the distribution of the manifest variables Y_i does not depend on the status of missingness given other observable test variables $\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}$ and/or observable covariates represented by $\boldsymbol{Z}$. Hence, *within* each subpopulation represented by the values $(\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}) = (\boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, z)$ the occurence of missing data with respect to Y_i is MCAR. As Equation 2.49 reveals, for each value $(\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}) = (\boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, z)$ the true scores do on average not differ between randomly drawn test takers who answer item i and those that produce an item nonresponse. The same holds true for each value $\boldsymbol{Z} = z$ if the missing data mechanism with respect to Y_i is MAR given $\boldsymbol{Z}$ (see Equation 2.51). Equivalently, for each value $\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}$ the true scores do on average not differ depending on responding to item i or not.

Implications of NMAR w. r. t True Scores If the missing data mechanism is NMAR missingness is called informative. In fact it can be shown that not only the distribution of Y_i varies depending on the observational status of item i but the underlying true score as well. From Equation 2.14 it follows that the probability of non-response with respect to item Y_i depends stochastically on Y_i even if all observable variables $(\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z})$ are held

constant statistically. Possibly, there are further unobserved variables such as motivation to take the test that are related to the performance on the test as well as to the non-response process of Y_i. Given all these variables , D_i and Y_i would be conditionally stochastically independent. However, if these variables are not observable in application, the missing data mechanism is NMAR. In this case, Equation 2.14 implies $\exists(\boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, z) \in \Omega_{\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}} \times \Omega_{\boldsymbol{Z}}$

$$g(Y_i \,|\, D_i = 1, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z) \neq g(Y_i \,|\, D_i = 0, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z). \tag{2.54}$$

Hence, in application the observed responses or values y_i of Y_i originate from a distribution that is different from the conditional distribution of the unobservable Y_i. Consequently, sample based inference is potentially biased due to data that are not representative with respect to the distributions of the manifest variables Y_i. Consequences of unconditional dependence $Y_i \not\perp D_i$ with respect to the true score distribution of τ_i were already examined in the previous section (see Equations 2.42 - 2.47). In the case of a nonignorable missing data mechanism, even in subpopulations given by the values $(\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}) = (\boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, z)$, the true score distributions will likely differ depending on the observational status. Let Y_i be a continuous normally distributed random variable with $N[0, Var(Y_i \,|\, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z)]$. If conditional independence $\varepsilon_i \perp D_i \,|\, (\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z})$ holds true then Equation 2.54 implies $\exists(\boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, z) \in \Omega_{\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}} \times \Omega_{\boldsymbol{Z}}$

$$g(\tau_i \,|\, D_i = 1, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z) \neq g(\tau_i \,|\, D_i = 0, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z} = z) \tag{2.55}$$

In the case of binary manifest variables Y_i from Equation 2.54, it follows immediately that Equation 2.55 holds as well, without additional assumptions with respect to ε_i. Hence, even conditional on observable variables, test takers that tend to respond to item i differ with respect to their true scores compared to those who tend to omit item i. When looking at real data, typically, persons with on average lower proportions of correct answers and, therefore lower true scores tend to omit items. As a consequence, in all subpopulations formed by the values of $(\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z})$ the observed responses to item i do not reflect the true average performance on this item or sub-test. Inference with respect to Y_i will not hold unconditionally and conditionally on $(\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z})$. The information in D_i needs to be taken into account to yield unbiased and consistent parameter estimates and valid sample based inference.

Implications with respect to constructed latent variables Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)$ be a M-multidimensional latent variable constructed in the measurement model constituted by

the items $Y_1, \ldots, Y_I$. In CTT and IRT, the variable $\boldsymbol{\xi}$ is defined as a function $\boldsymbol{\xi} = f(U)$ of the unit variable U. It is important to note that the true score variables are also latent variables in the sense that they are unobservable. In fact, as Borsboom and Mellenbergh (2002) stated, the latent variables $\xi_1, \ldots, \xi_M$ and the true scores are strongly related and in some measurement models they are equal. For example, in the case of a unidimensional linear SEM for parallel tests with factor loading equal to one and measurement intercepts equal to zero, it follows for all true score variables that $\tau_i = \xi$. However, in IRT models the conditional category probabilities $P(Y = y \,|\, \boldsymbol{\xi})$ are non-linear functions of latent variables $\xi_1, \ldots, \xi_M$ and not the true scores[2]. Generally, true scores are functions $\tau_i = f(\boldsymbol{\xi})$ of the latent variable. In the most general case $f()$ is any parametric or non-parametric function. In models of CTT and parametric IRT models, $f()$ is a parametric function whose parameters (item parameters) are aimed to be estimated and tested in application. In SEM these are the factor loadings and measurement intercepts, whereas in IRT models item difficulties or thresholds and the item discriminations are of interest. Since the true scores are functions of the latent variables, it seems straight forward to conclude that distributional differences of true scores between two or more conditions imply different distributions of latent variables. However, this conclusion is only valid if the function $f()$ is invariant across these compared conditions. In other words there is measurement invariance (Mellenbergh, 1989; Meredith, 1993; Lubke & Muthén, 2004) assumed with respect to $f()$. Meredith (1993) defined measurement invariance on the basis of the conditional distribution of the observed variables Y_i given particular covariates. Therefore, Y_i is measurement invariant with respect to $\boldsymbol{Z}$ given that

$$Y_i \perp \boldsymbol{Z} \,|\, \boldsymbol{\xi}. \tag{2.56}$$

This means that the observed score distribution depends exclusively on the distribution of the latent variable $\boldsymbol{\xi}$. Conditional on $\boldsymbol{\xi}$, Y_i is not stochastically dependent on the covariate $\boldsymbol{Z}$. Measurement invariance is always defined with respect to particular conditioning variables. Therefore it is possible that Y_i is measurement invariant with respect to $\boldsymbol{Z}$ but measurement invariance might not hold with respect to other covariates, for example $\boldsymbol{W}$. In the context of missing data methods, the response indicator variables D_i can be considered covariates too. Returning to the example of a SEM for τ-congeneric variables with a single latent variable ξ, suppose that $g(Y_i \,|\, D_i = 1) \neq g(Y_i \,|\, D_i = 0)$. Hence, the missing data mechanism with respect to Y_i is MAR or even NMAR. Furthermore $\varepsilon_i \perp D_i$

[2]Only in the case of dichotomous variables Y_i, monotonicity of item characteristic curves and simple structure $\xi_m = f(\tau_i)$ hold true, with f the link function.

holds true in this example implying $g(\tau_i \mid D_i = 1) \neq g(\tau_i \mid D_i = 0)$. To conclude that $g(\xi \mid D_i = 1) \neq g(\xi \mid D_i = 0)$ requires, however, that the factor loading and intercept are measurement invariant with respect to D_i.

In the remainder of this work the assumption of measurement invariance with respect to $(\boldsymbol{Z}, D_i)$ is assumed. That is

$$\forall i \in \{1, \ldots, I\} \;\; Y_i \perp (\boldsymbol{Z}, D_i) \mid \boldsymbol{\xi}. \tag{2.57}$$

This also implies measurement invariance with respect to D_i alone:

$$\forall i \in \{1, \ldots, I\} \;\; Y_i \perp D_i \mid \boldsymbol{\xi} \tag{2.58}$$

Additionally, we assume local stochastic independence for all manifest variables Y_i, that is

$$\forall i \in \{1, \ldots, I\} \;\; Y_i \perp \boldsymbol{Y}^{-i} \mid \boldsymbol{\xi} \tag{2.59}$$

Finally, we assume that Y_i is conditionally stochastically independent from $(\boldsymbol{Z}, D_i, \boldsymbol{Y}^{-i})$ given $\boldsymbol{\xi}$:

$$\forall i \in \{1, \ldots, I\} \;\; Y_i \perp (\boldsymbol{Y}^{-i}, \boldsymbol{Z}, \boldsymbol{D}) \mid \boldsymbol{\xi} \tag{2.60}$$

Note that Equation 2.60 follows neither from measurement invariance with respect to $(\boldsymbol{Z}, D_i)$ (see Equations 2.57) nor from local stochastic independence (see Equations 2.59). Conversely, however, if Equation 2.60 holds then the Equations 2.57 - 2.59 will apply as well.

Note that Equations 2.46 and 2.47 imply that the distribution $g(Y_i)$ is completely determined by the distribution $g(\tau_i)$ of the true scores. This applies also in the conditional case; The conditional distribution $g(Y_i \mid \boldsymbol{W})$ given any variable $\boldsymbol{W}$ is determined by the conditional true score distribution $g(\tau_i \mid \boldsymbol{W})$. The true score τ_i is the function $f_i(\boldsymbol{\xi})$ implying that the distribution of the manifest items Y_i is a composition $g(Y_i) = g[f_i(\boldsymbol{\xi})]$. In the conditional case that is $g(Y_i \mid \boldsymbol{W}) = g[f_i(\boldsymbol{\xi}) \mid \boldsymbol{W}]$. Measurement invariance means that $f_i(\boldsymbol{\xi})$ is an invariant function over all values $\boldsymbol{W} = \boldsymbol{w}$. In this case differences of conditional distributions of manifest items reflects necessarily differences of conditional distributions of true scores and the latent variable $\boldsymbol{\xi}$.

Hence, if the assumptions expressed by Equations 2.57 - 2.60 hold true the definitions of the missing data mechanisms with respect to the items Y_i imply (un-)conditional

stochastic dependencies between $\boldsymbol{\xi}$ and the response indicators and covariates:

- If the missing data mechanism w.r.t. Y_i is MCAR from Equation 2.44 follows

$$\boldsymbol{\xi} \perp D_i. \tag{2.61}$$

  Hence, the population of test takers that completed item Y_i differs in their distribution of the latent ability compared to those that did not complete it. However, given the missing data mechanism with respect to Y_i is MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$, then Equation 2.48 implies

$$\boldsymbol{\xi} \not\perp D_i. \tag{2.62}$$

  Hence, the population of test takers that completed item Y_i differs in their distribution of the latent ability compared to those that do not complete it. However, given the missing data mechanism with respect to Y_i is MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$, then Equation 2.48 implies

$$\boldsymbol{\xi} \perp D_i \,|\, (\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}). \tag{2.63}$$

  This implies that, although unconditional stochastic dependence between missingness and the latent ability holds (see Equation 2.62), test takers with the same values of observable variables $(\boldsymbol{Y}_{\boldsymbol{obs}}, \boldsymbol{Z})$ does not differ in their latent ability $\boldsymbol{\xi}$ regardless of whether responding to item Y_i or not.

- Similarly, if the missing data mechanism with respect to Y_i is MAR given $(\boldsymbol{Z})$, although Equation 2.62 applies, from Equation 2.12 follows

$$\boldsymbol{\xi} \perp D_i \,|\, \boldsymbol{Z}. \tag{2.64}$$

- Given the missing data mechanism with respect to Y_i is MAR given $\boldsymbol{Y}$ Equation 2.62 holds as well. However, the Equation 2.52 implies

$$\boldsymbol{\xi} \perp D_i \,|\, \boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}. \tag{2.65}$$

- If the nonresponse mechanism w.r.t. Y_i is NMAR the conditional distribution of the latent ability given all observable variables depends on the observational status

(D_i). From the definition (see Equation 2.14) and Equation 2.54 follow

$$\boldsymbol{\xi} \not\perp D_i \,|\, (\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}, \boldsymbol{Z}). \tag{2.66}$$

Thus, test takers who respond to Y_i differs systematically in their underlying ability levels from those who do not complete item i even if all observable variables are held constant.

Recall that all these implications hold under the assumptions of measurement invariance and local stochastic independence (see Equations 2.57 - 2.59). In the case of dichotomous items the implications given by the Equations 2.61 - 2.66 do not require additional assumptions with respect to the residual ε_i[3].

The assumption of measurement invariance is often made implicitly and seems reasonable to hold true in application. It seems not obvious why missingness should be related to parameters of the measurement model. However, examples can be constructed that make the assumption of measurement invariance unlikely to hold. For instance, let there be a mathematics test with a latent variable ξ representing mathematics proficiency. Assume that the last item Y_I is a mathematical problem formulated in text form. Additionally, a constructed response needs to be given by test takers. Some of the examinees might have a mother tongue different from the language used in the test. Therefore, they are on average slower in completing the items and, therefore, more likely not to reach the last item. Furthermore, they have on average a lower probability to solve Y_I. Let Z be the covariate indicating whether test takers' mother tongues are equal to the language of the test ($Z = 1$) or not ($Z = 0$). This example implies that the probability of missing item Y_I is $P(D_I = 1 \,|\, Z = 1) > P(D_I = 1 \,|\, Z = 0)$. Additionally, we assumed $P(Y_I = 1 \,|\, Z = 1) > P(Y_I = 1 \,|\, Z = 0)$. The missing data mechanism w.r.t. Y_I is assumed to be MAR given Z, that is $D_I \perp Y_I \,|\, Z$. As a consequence, the item Y_I will be more frequently answered by persons with a mother tongue equal to the language of the test. These persons have also a higher probability to answer correctly. Assuming that persons with a mother tongue different from the language of the test have on average the same mathematical ability, the lower probabilities to solve item I can be attributable to differential item functioning (DIF) with respect to Z. Hence $P(Y_I = 1 \,|\, \xi, Z) \neq P(Y_I = 1 \,|\, \xi)$. This might be due to the demanding text. As a results the assumption of conditional stochastic independence $Y_I \perp (D_I, Z) \,|\, \xi$ (see Equation 2.57) is violated in this example. This has

[3]For normally distributed manifest variables Y_i with linear functions $f_i(\boldsymbol{\xi})$, additional assumptions with respect to ε_i are required.

interesting consequences: Let $P(Y_I = 1 \,|\, \xi, Z, D_I) = P(Y_I = 1 \,|\, \xi, Z)$. Since, Z and D_I are stochastically dependent in this example, it follows $P(Y_I = 1 \,|\, \xi, D_I) \neq P(Y_I = 1 \,|\, \xi)$. Hence, if a covariate Z exists that causes DIF and Z is also stochastically related to the probability of non-response, then measurement invariance with respect to D_I is unlikely to hold. DIF is a common phenomenon as well as missing data. Hence, the short example used for illustration seems not too unrealistic and should make aware that the assumption of measurement invariance with respect to D_i and $(D_i, \boldsymbol{Z})$ can be violated.

2.4 Summary

In this section the different missing data mechanisms were defined. Instead of three, five different nonresponse mechanisms are distinguished (a) Missing completely are random (MCAR), (b) missing at random given $(\boldsymbol{Y}, \boldsymbol{Z})$, (c) missing at random given $\boldsymbol{Z}$, (d) missing at random given $\boldsymbol{Y}$, and (e) missing not at random (MNAR). The differentiation into three MAR conditions result from the distinction between manifest variables $\boldsymbol{Y} = Y_1, \ldots, Y_I$ that constitute the measurement model and covariates $\boldsymbol{Z} = Z_1, \ldots, Z_J$. In this work it is assumed that the covariates are fully observed in application. The nonresponse mechanisms were defined with respect to single items Y_i and, subsequently with respect to the complete response vector $\boldsymbol{Y}$. Following Rubin $\boldsymbol{Y}$ is decomposed in an observed part $\boldsymbol{Y_{obs}}$ and an unobserved part $\boldsymbol{Y_{mis}}$. In contrast to Rubin, missing data mechanisms were defined here using random variables considered in a particular random experiment - the single unit trial - instead of realized data. Hence, the definitions rest upon the joint distribution $g(\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}, \boldsymbol{Z}, \boldsymbol{D})$. As Kenward and Molenberghs (1998) noted, the missing data mechanisms as introduced in most statistical literature seems to be confusing for non-bayesian statisticians and methodologists. Due to the adaption of the definitions in this work, consistency with the pre-facto perspective and frequentists' estimation theory, such as ML estimation, has been achieved. Nevertheless, the essentials of Rubin's definitions are preserved. It was shown in detail that data matrices resulting under the re-defined missing data mechanisms will have the properties described by Rubin if the sample size becomes large. Therefore, the definitions presented here have been formally adapted for reasons of consistency but are in accordance with existing missing data literature.

In the final section of this Chapter the implications of the nonresponse mechanisms regarding the latent variables underlying observed and unobserved data were examined theoretically. It was explained why ignorable missing data mechanisms are called noninformative, whereas informative missingness refers to nonignorable missing data. It was

shown analytically that the occurence of item nonresponses are completely at random if MCAR w.r.t. $\boldsymbol{Y}$ holds true, and completely at random *conditionally* on all possible values $(z, \boldsymbol{y}_{obs})$ if MAR w . r. t. $\boldsymbol{Y}$ applies. Hence, apart from a loss of efficiency inference with respect to parameters ι will be unbiased in the complete sample (MCAR) or in subsamples formed by observed values $(z, \boldsymbol{y}_{obs})$ (MAR). In the latter case the information needs to be appropriately aggregated including all observable variables across all missing patterns. In fact, FIML can be regarded as aggregating information over all observed values $(z, \boldsymbol{y}_{obs})$ in all observed missing pattern $\boldsymbol{D} = \boldsymbol{d}$. Accordingly, missingness expressed by $\boldsymbol{D}$ does not provide additional information with respect to parameters of interest and can, therefore, be ignored in sample based inference. This was also illustrated considering multiple imputation. Finally, it was shown that under particular assumptions of conditional stochastic independence (see Equations 2.57 - 2.60) the distributions of the observed and unobserved manifest variables differ, which implies that populations of the test takers who complete an item compared to those who do not differs with respect to the latent variable $\boldsymbol{\xi}$ of interest. This is the case when the missing data mechanism w.r.t. Y_i is MAR or NMAR. However, when one of the MAR conditions hold w.r.t. $\boldsymbol{Y}$ (see Equations 2.10 - 2.13), the distribution of $\boldsymbol{\xi}$ underlying observed and unobserved manifest items Y_i are conditionally equal given each value $(z, \boldsymbol{y}_{obs})$. This is not true when the missing data mechanisms are nonignorable. What does this mean for applied research? In the subsequent section the effects of non-ignorable missing data to sample based inference will be examined. In general it should be noted that in measurement models many manifest variables $Y_1, \ldots, Y_I$ are considered contemporarily. Each item can be affected by a different missing data mechanism. As a consequence each item is potentially completed by a different population even in a single test application. The missing data mechanism works as an item specific selection mechanism.

3 The Impact of Missing Data on Sample Estimates

Missing data might affect sample based inference in many different ways. A general description of the impact of missing data is difficult. Validity and accuracy of inference under missing data might be affected differently depending on the particular research question, the data, the missing data mechanism, and the applied models. That is an important reason why the problem of missingness has been studied separately in different contexts and why specific approaches need to be developed. Of course these methods and approaches can roughly be classified (e. g. Schafer & Graham, 2002; McKnight et al., 2007; Graham, 2009). A brief overview is given in section 4.1 in order to integrate the methods examined in this work. However, before the approaches tackling the problem of missing data in measurement models will be examined in detail, the impact of missing data will be illustrated. The focus is on non-ignorable missing data due to nonresponses . However, the derivations and results will be repeatedly linked to cases where the non-response mechanism is MAR or even MCAR. Nonresponses in educational and psychological testings can result from omitting items, providing answers that are not meaningful and therefore not codable, or not reached items at the end of the test. This work only marginally deals with unit-nonresponses. That does not mean that it is not a serious problem in real applications. Therefore, this work deals with incomplete data sets and how to account for the problems associated with them. However, many of the illustrated problems due to item nonresponses are close to those caused by unit nonresponses.

In this chapter, the impact of missing data will be studied with respect to person and item parameter estimates. There are different measures to describe the items with respect to their difficulty and discriminating power. Analogously, several measures or person parameters exist to quantify persons' achievement in the test and/or to locate test takers with respect to the latent variable constructed in the measurement model. The measures can be classified into two groups. Most psychometrically developed tests are based on Classical Test Theory (CCT) or Item Response Theory (IRT). The person parameters in CTT rest upon (un-)weighted sum scores or (non-)linear functions of it. The difficulties of items are expressed by the item means. Point-biserial and biserial correlations between single

items and the test score serves as discrimination parameters. In IRT item parameters and their meaning depend on the respective model chosen in a particular application. In the 1PL- and 2PL models, the most frequently used IRT models, the item discrimination parameter is equivalent to a logistic or probit regression coefficient and the item difficulty is a transformed logistic or probit regression intercept. Person parameter estimates are direct estimates of the persons' individual values on the latent variable constructed in the measurement model. CTT and IRT are quite different test theories (Embretson & Reise, 2000; Fan, 1998; Hambleton & Jones, 1993)CTT focus more on the test-score level than on individual items (Hambleton & Jones, 1993) and measurement models of CTT are inappropriate for dichotomous items. Nevertheless, the study of the impact of missing data regarding to CTT-based item and person parameter estimates in tests with binary items are valuable in understanding harmful effects of item nonresponses. In this thesis, the effects of missing data will be separately studied for CTT and IRT item and person parameter estimates. The considerations will comprise analytical derivations and empirical illustrations by simulated data examples. There are two reasons for the use of simulations. At first, the impact of missingness can be studied and quantified under varying conditions. Secondly, for some of the parameter estimates no closed-form expressions exist for the respective estimation equation. Hence, the bias is difficult to determine analytically. Single model parameters needs to be estimated iteratively depending on other unknown model parameters that are estimated contemporarily. This is in particular true for the IRT models, where estimates of item difficulties and discriminations are mutually dependent. At least hypotheses can be formulated about the expected bias due to the nonresponse mechanism that can be supported or falsified by the simulated data. However, the impact of missing data on sample estimates is studied analytically as far as possible.

A test typically consists of a set of stimuli, the items, that elicit a response behavior in test taker. The item responses are indicative of the latent variable which is constructed in the measurement model. A sound and well-founded test development comprises the quantification of the quality of psychometric properties of the test with respect to certain quality criteria. The objectivity, the reliability, and the validity are the so-called main quality criteria (Amelang & Zielinski, 2001). Additionally, a considerable number of further quality criteria range from the theoretical foundation of the test construction to the layout of the test and its manual (Amelang & Zielinski, 2001). Of course, missing data might also influence the measures and indices used to quantify the psychometric quality of a test. Here it is impossible to study all potential effects. The considerations are confined to the impact of item-nonresponses on reliability and test fairness, knowing full well that the

whole range of adverse effects is not covered. Different measures of reliability exist and can also be attributed to the two major classes of test theories. Specifically, Cronbach's α (Cronbach, 1951) and Guttmann's λ_2 (Guttman, 1945) are widely used in CTT based test development. These coefficients are suited when the manifest test variables are linearly regressively dependent from the latent variable. For example, Cronbach's α is appropriate under the model of essentially τ-equivalent variables (Steyer & Eid, 2001). However, this work focuses on dichotomous manifest variables. The regression $E(Y_i \mid \boldsymbol{\xi})$ considered in the measurement model with categorical manifest variables is almost never linear. Hence, Cronbach's α and Guttmann's λ_2 are unsuitable coefficients of reliability and will not be considered in this thesis. In IRT it is common to utilize the item information function and/or the standard error function to describe the accuracy and/or the error of the person parameter estimation. Insofar, IRT accounts for the fact that a test might be more or less accurate for different test takers depending on their values of the latent variable. This implies that the reliability varies accross the range of the latent variable. Nevertheless, summary measures of reliability have been developed in IRT and are widely used. Typically, Andrich's reliability (Andrich, 1988) or the EAP-reliability (Bock & Mislevy, 1982) are used. Both can be interpreted as mean reliability coefficients averaged across the distribution of the latent variable. If the maximum likelihood estimators (MLE) or Warm's weighted maximum likelihood estimators (WLE) are used, Andrich's reliability is appropriate whereas the EAP-reliability is taken if the expected a posteriori (EAP) estimators are chosen as person parameter estimates. The reliability coefficients are determined based on item parameters and person parameters and its distribution. Thus, the reliability might be affected in different ways due to nonresponses: Firstly, due to missing information, and secondly, because of biased parameter estimates.

To sum up, in this chapter the impact of missing data on CTT-based and IRT-based item and person parameter estimates are studied analytically and empirically. A considerable number of different IRT-models have evolved. Here, only the one-parameter Rasch-Model (1PLM) and the two-parameter Birnbaum-Model (2PLM) will be considered. The marginal reliability coefficients with respect to MLE, WLE, and EAP estimators are examined under different missing data situations. Against the background of these results the matter of test fairness in presence of missing data will be critically discussed.

As previously mentioned, the theoretical examination of the bias due to missing data is limited in some cases. The illustration of the effects of missing data with simulated data is based on (a) a single simulated data set suffering from a high proportion of non-ignorable missing data, and (b) a comprehensive simulation study with varying conditions.

The simulated Data Example A with non-ignorable missing data The data set introduced here will be used in the remainder of this thesis to demonstrate the harmful effects due to ignoring missing data or the application of inappropriate missing data methods. Furthermore, the suitability of the proposed methods for non-ignorable missing data will be exemplified with this data set denoted by Data Example A in the remainder. The application of a test consisting of $I = 30$ dichotomous items Y_i was emulated. Hence, the simulated data can be thought of as resulting from an application of a reading or mathematics achievement test with the response category $Y_i = 0$ indicating a wrong answer and $Y_i = 1$ the correct answer. The sample size was $N = 2000$. The latent ability variable ξ was unit normally distributed with $E(\xi) = 0$ and $Var(\xi) = 1$. The item responses where simulated using the 1PLM:

$$P(Y_i = 1 \,|\, \xi) \quad = \quad \frac{\exp(\xi - \beta_i)}{1 + \exp(\xi - \beta_i)} \tag{3.1}$$

The item difficulties are equally spaced between -2.3 and 2.15. The difference between two subsequent difficulties is 0.15. The probability of nonresponses was stochastically related to the latent variable ξ. In the realized data the sample correlation between the latent variable ξ and the proportion of missing data was $r = -0.719$. The data were simulated in that way such that the probability to omit items increases with lower values of ξ. This emulates the often reported finding that the incidence of non-responses increases with declining proficiency levels. Possibly, less proficient persons tend to respond to items they judge to solve correctly. Furthermore, difficult items may require more cognitive efforts especially for test takers with lower ability levels. Especially in low stakes assessments, test takers might not be motivated and/or unwilling to make such efforts. This increases also the probability of missing data with decreasing ability levels. Finally, the processing time with respect to single items may prolonged with decreasing values of ξ resulting in missing data due to not-reached items at the end of the test.

For all items in data example $P(D_i = 1) < 1$ holds. The missing data mechanism with respect to each item Y_i was NMAR. Accordingly the nonresponse mechanism w.r.t. $\boldsymbol{Y}$ is nonignorable (see Section 2.2). The individual probability $P(D_i = 1 \,|\, U = u)$ to respond to item i was obtained by the introduction of a latent response propensity $\theta = f(U)$ as a function of the person variable U. θ can be thought of as a tendency of the test takers to complete the test items. The specific item response propensities $P(D_i = 1 \,|\, U = u) = P(D_i = 1 \,|\, \theta)$ are a function of the latent variable θ. The probability to respond to item Y_i,

regardless of whether correctly or incorrectly, is given by

$$P(D_i = 1 \mid \theta) \quad = \quad \frac{\exp(\theta - \gamma_i)}{1 + \exp(\theta - \gamma_i)} \tag{3.2}$$

This equation is equivalent to the 1PLM. The parameters γ_i are the thresholds of the respective response indicator variables D_i. In the data example the parameters γ_i ranges between -2.57 and 2.06. The data are generated under conditional stochastic independence $D_i \perp \xi \mid \theta$ and $Y_i \perp \theta \mid \xi$. Hence, in Data Example A, non-ignorability of the missing data mechanism is implied by the correlation $Cor(\xi, \theta) = 0.8$. For the single items Y_i Equation 2.14 holds. Thus, if a measurement model is exclusively estimated based on $\boldsymbol{Y}$ item, then person parameters are potentially biased. In real applications of achievement tests, it is a consistent finding, that more difficult items are generally more often skipped (Rose et al., 2010). This implies that the parameters γ_i and β_i are also related. For the simulated data example, this dependency is presented graphically in Fig. 3.1. The higher the values of β_i are, the higher γ_i is. The means of the probabilities $P(Y_i = 1 \mid \xi)$ and

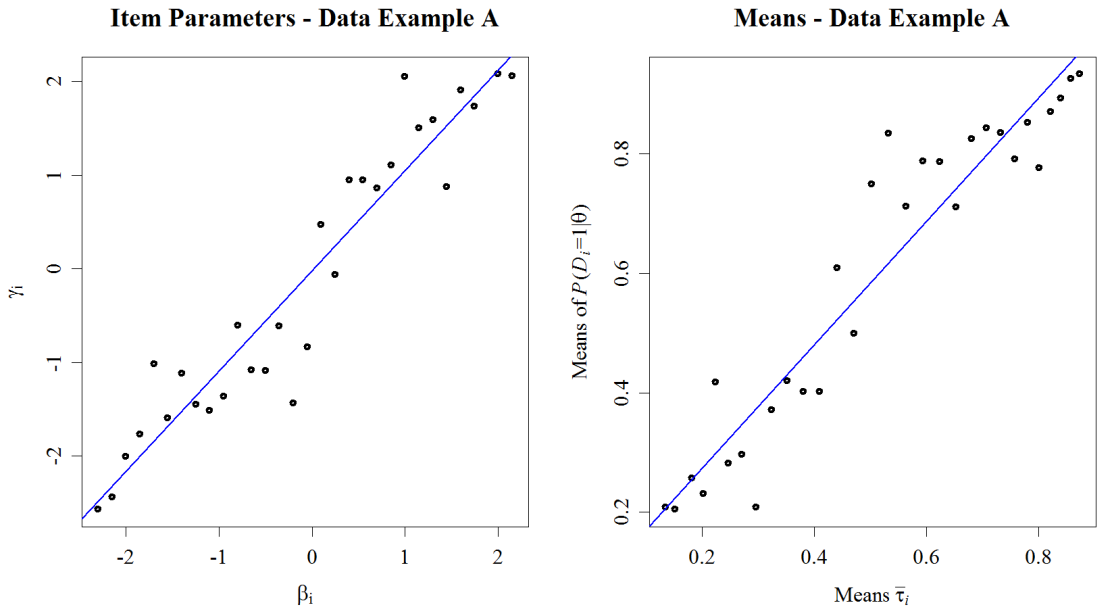

Figure 3.1: Item difficulties and thresholds used to generate Data Example A (left) and resulting means $\bar{\tau}_i$ of true scores and item response propensities (right). The blue line is the regression line.

$P(D_i = 1 \mid \theta)$ are plotted for each item i in the right panel of Figure 3.1. It can be seen that the most difficult items are merely expected to be completed by $\approx 20\%$. The overall proportion of missing data in the realized data was 47.83%. Across the items it ranges

between 80.2% and 6.6% (Tab. 3.1). Compared with real applications, the conditions

Table 3.1: Item Parameters of Items Y_i, Response Indicators D_i and Marginal Probabilities $P(Y_i = 1)$ and $P(D_i = 1)$ (Data Example A).

Items			Response indicators		
Y_i	β_i	$P(Y_i = 1)$	D_i	γ_i	$P(D_i = 1)$
Y_1	-2.30	0.872	D_1	-2.569	0.934
Y_2	-2.15	0.857	D_2	-2.440	0.926
Y_3	-2.00	0.840	D_3	-2.009	0.893
Y_4	-1.85	0.821	D_4	-1.768	0.871
Y_5	-1.70	0.801	D_5	-1.016	0.777
Y_6	-1.55	0.780	D_6	-1.597	0.853
Y_7	-1.40	0.757	D_7	-1.118	0.792
Y_8	-1.25	0.733	D_8	-1.450	0.836
Y_9	-1.10	0.707	D_9	-1.517	0.844
Y_{10}	-0.95	0.680	D_{10}	-1.363	0.825
Y_{11}	-0.80	0.652	D_{11}	-0.608	0.711
Y_{12}	-0.65	0.623	D_{12}	-1.081	0.787
Y_{13}	-0.50	0.594	D_{13}	-1.091	0.788
Y_{14}	-0.35	0.563	D_{14}	-0.615	0.712
Y_{15}	-0.20	0.533	D_{15}	-1.437	0.834
Y_{16}	-0.05	0.502	D_{16}	-0.838	0.749
Y_{17}	0.10	0.471	D_{17}	0.474	0.499
Y_{18}	0.25	0.440	D_{18}	-0.066	0.609
Y_{19}	0.40	0.410	D_{19}	0.950	0.402
Y_{20}	0.55	0.380	D_{20}	0.950	0.401
Y_{21}	0.70	0.351	D_{21}	0.861	0.419
Y_{22}	0.85	0.323	D_{22}	1.106	0.371
Y_{23}	1.00	0.297	D_{23}	2.054	0.208
Y_{24}	1.15	0.271	D_{24}	1.503	0.297
Y_{25}	1.30	0.246	D_{25}	1.589	0.282
Y_{26}	1.45	0.223	D_{26}	0.872	0.417
Y_{27}	1.60	0.202	D_{27}	1.905	0.230
Y_{28}	1.75	0.182	D_{28}	1.736	0.257
Y_{29}	2.00	0.151	D_{29}	2.078	0.205
Y_{30}	2.15	0.135	D_{30}	2.059	0.208

used for Data Example A may be exaggerated. For example, the PISA 2006 data were re-analyzed by Rose et al. (2010). They modeled a latent response propensity allowing to estimate the correlation between ξ and θ. In the PISA study a three-dimensional latent variable $\boldsymbol{\xi}$ was constructed with the latent mathematic dimension ξ_{math},the latent

reading dimension ξ_{read}, and the latent science dimension ξ_{sci}. The correlations between these latent variables and the latent response propensity θ varied between $0.267 - 0.555$ accross the 30 member states of the Organization for Economic Co-operation and Development (OECD) considered in the study. The mean correlations were $\bar{r}(\xi_{math}, \theta) = 0.433$, $\bar{r}(\xi_{read}, \theta) = 0.434$ and $\bar{r}(\xi_{sci}, \theta) = 0.453$. Additionally, the means of the response indicators and the standardized item means were positively correlated with $r = 0.330$ in the PISA 2006 data. Hence, more difficult items were generally more often not answered than less difficult items. Compared with these results, the conditions used for Data Example A are accentuated for reasons of demonstrating the effects of non-ignorable missing data. In real situations, the conditions as the total proportion of missing data, the sample size, the number of items and the missing data mechanism might differ. Therefore, an additional simulation study was conducted in order to quantify the impact of missing data under different conditions.

The simulation study As noted previously, for some parameters, such as IRT item and person parameters, the impact of missing data is difficult to examine analytically. The simulation study aims to find general patterns of biases and to quantify the extent of biasedness of different parameter estimates caused by item nonresponses. This is particularly of interest, since results of single simulated data examples by Rose et al. (2010) might suggest that IRT parameter estimates are pretty robust under non-ignorable missing data.

In the simulation study five factors were systematically varied. Three sample sizes $N = \{500, 1000, 2000\}$ were chosen. The effect size of the relation between the probability of missing data and the latent variable ξ was controlled by different correlations $Cor(\xi, \theta) = \{0, 0.2, 0.5, 0.8\}$. Note, that the missing data mechanism is MCAR given that $Cor(\xi, \theta) = 0$. Three different test length were simulated. The numbers of items were 11, 22 and 33. The relation between the item parameters γ_i and β_i were varied as well with the approximate values $r(\gamma, \beta) = \{0, 0.3, 0.5, 0.8\}$. In this case, $r(\gamma, \beta)$ is computed in the same way as the sample estimate of the Pearson-correlation. Of course, the parameters γ_i and β_i are not realizations of random variables. However, the value $r(\gamma, \beta)$ is suited to quantify the relation between the difficulty of an item and the unconditional probability of non-response. In contrast, if $r(\gamma, \beta) > 0$ then $\beta_i > \beta_j \Rightarrow P(D_i = 1) < P(D_j = 1)$. Thus, the more difficult items are generally more likely to be not completed. Finally, the overall proportion of missing data was varied by different sets of parameters γ_i. A constant term was added to or subtracted from to all thresholds γ_i to yield average proportions of 10%,

20%, 30%, 40%, and 50% missing data. Since the parameters γ_i were only shifted by constants the correlation $r(\gamma, \beta)$ remained unaffected. Table 3.2 summarizes the factors and their levels used in the simulation study. In total there are 720 conditions in the

Table 3.2: Factors and Factor Levels Used for the Simulation Study.

Sample size N	$N = \{500, 1000, 2000\}$
Number of Items I	$I = \{11, 22, 33\}$
Correlation $Cor(\xi, \theta)$	$Cor(\xi, \theta) = \{0, 0.2, 0.5, 0.8\}$
Parameter correlation $r(\gamma, \beta)$	$r(\gamma, \beta) = \{0, 0.3, 0.5, 0.8\}$
Average proportion of missing data $\bar{P}_{Miss}$	$\bar{P}_{Miss} = \{10\%, 20\%, 30\%, 40\%, 50\%\}$

simulation study. 50 data sets were simulated within each condition.

The dependent variable in the simulation study was the biases of different IRT item and person parameter estimates. In particular, the bias of estimated item difficulties and item discrimination were studied as well as the bias of three person parameter estimates: Maximum Likelihood (ML) estimates, Warm's weighted Maximum Likelihood (WML) estimates, and EAP estimates. The bias of a parameter estimate $\hat{\lambda}$ is defined as the difference $\hat{\lambda} - \lambda$. In a simulation study with Q trials per condition the bias of a single parameter λ is

$$Bias(\lambda) = \frac{1}{Q} \sum_{q=1}^{Q} (\hat{\lambda}_q - \lambda). \tag{3.3}$$

However, in this simulation study there are I items and N person parameters. Hence, there is a vector $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_K$ of parameters. Accordingly, the mean bias $Bias(\boldsymbol{\lambda})$ was computed across all considered item or person parameter estimates. That is

$$Bias(\boldsymbol{\lambda}) = \frac{1}{Q} \sum_{q=1}^{Q} \Big[\frac{1}{K} \sum_{l=1}^{K} (\hat{\lambda}_{ql} - \lambda_l) \Big]. \tag{3.4}$$

If the person parameters are considered K is the sample size N an $\boldsymbol{\lambda} = \xi_1, \ldots, \xi_N$. For the case of item parameters K is the number I of items with $\boldsymbol{\lambda} = \alpha_1, \ldots, \alpha_I$ and $\boldsymbol{\lambda} = \beta_1, \ldots, \beta_I$ respectively.

The software *R* (R Development Core Team, 2011) was used for data simulation. The model parameters shown in Table 3.1 were used to simulate item responses and missing data. The data were generated using the Rasch model (Rasch, 1960) and the Birnbaum model (Birnbaum, 1968). In the latter, the item discrimination parameters were randomly

drawn from a continuous uniform distribution $U(0.5, 2.5)$. The parameters α_i varied from trial to trial but ranged always between 0.5 and 2.5. M*plus* 6 (Muthén & Muthén, 1998 - 2010) was used to estimated parameters of the Birnbaum model. Unfortunately, only EAPs are available in M*plus* as person parameter estimates. The effects of missing data on MLE and WLE person parameter estimates were studied under the 1PLM with ConQuest (Wu, Adams, & Wilson, 1998).

3.1 Test Scores and Person Parameter Estimates

Based on the observed response patterns of test takers, many different measures can be constructed that aim to quantify the persons' values of latent variables. Within the framework of CTT, the test scores are commonly based on the unweighted or weighted sum score S or functions of S. In order to obtain meaningful values, the sums score is usually standardized or tranformed in other ways. The standard scores (e. g. z-values, *Stanine*-values, T-values) are linear transformations of the sum score. Alternatively, non-linear transformations of S can be used such as the percent rank. However, in CTT the resulting test scores are only meaningful with respect to a particular population that serves as a reference. For example, a z-score of a test is the difference between a test score and the mean of the test scores in a certain population measured in standard deviations. CTT based testings are norm-referenced assessment. Alternatively, the proportion correct P^+ can be used instead of S. P^+ is the relative frequency of correctly answered items given the completed items instead of all presented items. If all persons answer the same items and no missing data exist, P^+ is simply a function of S. However, in presence of item-nonresponses S and P^+ are no longer deterministically related implying that both scores are differently affected by missing data. In fact, in application P^+ is typically preferred for incomplete data because the number of completed items is taken into account. However it needs to be answered whether P^+ is always sufficient to account for missing responses even when the nonresponse mechanism is MAR or NMAR. The impact of missing data to the sum score S and the proportion correct P^+ will be compared in Sections 3.1.1 and 3.1.2.

IRT based person parameters are differently constructed. Usually weighted or unweighted ML estimates or Bayesian estimates such as the EAP or maximum a posteriori (MAP) are used as measures of a person?s ability. Instead of the CTT-based test scores, the person parameters are values of the latent variable ξ and item difficulty parameters are located on the same scale. Persons and items can directly be compared in terms of

ability and difficulty. The values of the latent variables are also meaningful with respect to items. Conclusively, IRT-based measurement is a criterion-referenced assessment. In application typically neither item nor person parameters are known but need to be estimated based on observed data. Both item and person parameter estimates are mutually dependent. This is easy to see when Joint Maximum Likelihood (JML) estimation is considered where all item and person parameters are estimated simultaneously (e. g. Baker, 1987; Baker & Kim, 2004). Biased item parameter estimates will, most likely, result in biased person parameter estimates. In turn biased person parameter estimates can cause distorted item parameter estimates. Although less obvious, this also the case when using Marginal Maximum Likelihood (MML) estimation (e. g. Baker, 1987; Baker & Kim, 2004; Bock & Aitkin, 1981; Bock & Lieberman, 1970). Using MML item parameters are estimated separately avoiding simultaneous person parameter estimation[1]. However, using the EM algorithm (Bock & Aitkin, 1981; Hsu, 2000) item parameter estimation involves the calculation of probabilities $P(\boldsymbol{\xi_q} \mid \boldsymbol{Y} = \boldsymbol{y}; \hat{\boldsymbol{\iota}}_t)$ in the E-step for the evaluation of the quadrature distribution $g(\boldsymbol{\xi}_q \mid \boldsymbol{Y} = \boldsymbol{y}; \hat{\boldsymbol{\iota}}_t)$ of each test taker. $\boldsymbol{\xi_q}$ is the q-th quadrature point and $\hat{\boldsymbol{\iota}}_t$ the vector of estimated item parameters in the t-th iteration. Hence, although the point estimation of $\boldsymbol{\xi}$ is circumvented under MML estimation, the quadrature distribution of the latent variable $\boldsymbol{\xi}$ is still involved. Using MML, the person parameters are estimated in a second step with the estimated item parameters taken as fixed values. Due to the interdependence of IRT item and person parameter estimates the analytical examination of their bias due to item nonresponses is not feasible. For that reason, the bias of ML, WML, and EAP estimates are investigated by means of a simulation study. The results will be shown in Section 3.1.3.

3.1.1 Sum score

The sum score S is defined as the sum $S = \sum_{i=1}^{I} Y_i$ over all I items. For dichotomous items Y_i it is the number of correctly answered items. That is why S is sometimes called number right score. For theoretical reasons here it is distinguished between the sum score S in absence of missing data and the sum score S_{Miss} in presence of missing data. Although both S and S_{Miss} are number right scores, S_{Miss} is the sum across the completed items whereas S is the sum across all I items. Therefore, in presence of any previously defined missing data mechanism w.r.t. to the items Y_i the sum score S_{Miss} for a randomly chosen

[1] Individual person parameters can be estimated subsequently based on the previously estimated item parameter estimates.

observation is given by

$$S_{Miss} = \sum_{i=1}^{I} Y_{i|D_i=1}. \tag{3.5}$$

The condition $D_i = 1$ reflects that in application only those items can be summed which are observed. Hence, for each case the number of completed items is the upper bound of the sum score variable. Note, that the number of completed items varies across the test takers, the upper bound is itself a random variable given by $\sum_{i=1}^{I} D_i$. This fact is not taken into account when the sums score is used in real applications. Consider Data Example A, which consists of 30 items. In presence of missing data the score $S = 10$ is related to different events. For example, a test taker could have answered 10 items correctly while 20 items were answered incorrectly. Alternatively, a participant could have omitted 20 items but answered 10 items correctly. The sum score does not adequately account for non-responses. It can be shown that the sum score implicitly recodes missing responses as incorrect or more generally $Y_i = 0$ regardless whether the items are omitted, notreached, or even not presented by design. Formally, this can be represented by using the response indicators D_i as weights for Y_i. The sum score S_{Miss} as defined above can alternatively be written as

$$S_{Miss} = \sum_{i=1}^{I} Y_i \cdot D_i. \tag{3.6}$$

In this Equation, the sum is taken over all I items of the test. The sum score S_{Miss} is then a sum of a product variable $Y_i \cdot D_i$. The value of this variable is computed over all I items. Each term $Y_i \cdot D_i$ becomes zero if either the item Y_i is answered incorrectly or the item is not completed. In many applications, especially in educational large scale assessments non-responses are treated as wrong responses by assigning the value 0. Formally, a random variable Y_i^* can be defined as a function $f(Y_i, D_i)$ that is given by the following assignment rule:

$$Y_i^* = \begin{cases} Y_i, & \text{if } D_i = 1 \\ 0, & \text{if } D_i = 0 \end{cases} \tag{3.7}$$

Interestingly, the product variable $Y_i \cdot D_i$ and Y_i^* are equal proving that the use of the sum score under any missing data mechanism means to recode missing data to wrong

responses implicitly[2]. So, S_{Miss} is considered instead of S. However, S_{Miss} is the sum of $Y_i \cdot D_i$ or Y_i^* respectively instead of items Y_i. Summing over different random variables results in different sum scores with different distributions and potentially a different meaning. For the case that each test taker has a positive probability to answer missing items correctly, the sum score is expected to be negatively biased.

In order to study how non-ignorable missing data affects the sum score, the expectation of S_{Miss} is considered that can be written as

$$\begin{aligned} E(S_{Miss}) &= E\Big(\sum_{i=1}^{I} Y_i \cdot D_i\Big) && (3.8)\\ &= \sum_{i=1}^{I} E(Y_i \cdot D_i) && (3.9)\\ &= \sum_{i=1}^{I} E[E(Y_i \cdot D_i \,|\, U)] && (3.10) \end{aligned}$$

Equation 3.10 shows that expected value of each product variable $Y_i \cdot D_i$ is the expectation of the regression $E(Y_i \cdot D_i \,|\, U)$ studied next. The regression of a product variable is given by:

$$E(Y_i \cdot D_i \,|\, U) = E(Y_i \,|\, U) \cdot E(D_i \,|\, U) + Cov(Y_i, D_i \,|\, U) \tag{3.11}$$

The last summand is the conditional covariance that can be written as:

$$\begin{aligned} Cov(Y_i, D_i \,|\, U) &= E\Big([Y_i - P(Y_i = 1 \,|\, U)] \cdot [D_i - P(D_i = 1 \,|\, U)] \,|\, U\Big) && (3.12)\\ &= E(\varepsilon_{Y_i} \cdot \varepsilon_{D_i} \,|\, U) && (3.13)\\ &= Cov(\varepsilon_{Y_i}, \varepsilon_{D_i} \,|\, U) && (3.14) \end{aligned}$$

In the subsequent derivations it is assumed that the conditional covariance $Cov(\varepsilon_{Y_i}, \varepsilon_{D_i} \,|\, U)$ is zero. Hence, for dichotomous variables Y_i Equation 3.20 can be simplified to

$$\begin{aligned} E(Y_i \cdot D_i \,|\, U) &= E(Y_i \,|\, U) \cdot E(D_i \,|\, U) && (3.15)\\ &= P(Y_i = 1 \,|\, U) \cdot P(D_i = 1 \,|\, U). && (3.16) \end{aligned}$$

[2]Note that this statement is only valid if $Y_i = 0$ indicates a wrong response. For example SAT scoring is different. $Y_i = -0.25$ indicates a wrong response and $Y_i = 0$. Under such a scoring missing responses are not implicitly recoded to an incorrect answer.

The term $E(Y_i \cdot D_i \mid U)$ can also be expressed as the conditional probability $P(Y_i = 1 \cap D_i = 1 \mid U)$. Thus, it can be seen that the assumption $Cov(\varepsilon_{Y_i}, \varepsilon_{D_i} \mid U)$ is equivalent to the assumption of conditional stochastic independence of $Y_i = 1$ and $D_i = 1$ given the person variable U.
Utilizing these derivations, we can consider the conditional expected sum score $E(S_{Miss} \mid U)$ given any missing data mechanism under the assumption $Y_i \perp D_i \mid U$.

$$
\begin{aligned}
E(S_{Miss} \mid U) &= E\Big(\sum_{i=1}^{I} Y_i \cdot D_i \,\Big|\, U\Big) && (3.17)\\
&= \sum_{i=1}^{I} E(Y_i \cdot D_i \mid U) && (3.18)\\
&= \sum_{i=1}^{I} P(Y_i = 1 \mid U) \cdot P(D_i = 1 \mid U) && (3.19)
\end{aligned}
$$

Here, it can directly be seen that the expected S_{Miss} given the person projection U is smaller under any missing data mechanism compared to the expected sum score $E(S \mid U)$. Only if no missing data mechanism exist, so that $P(D_i = 1 \mid U) = 1$ (for all $I = 1, \ldots, I$), equality $E(S_{Miss} \mid U) = E(S \mid U)$ follows. The difference $S_{Miss} - S$ can be regarded as a bias of the sum score resulting form missing data. Since $Y_i \cdot D_i \leq Y_i$, the bias can never be positive. The expected conditional bias $E(S_{Miss} - S \mid U)$ given the unit variable U can be studied in more detail, starting with the following Equations.

$$
\begin{aligned}
E(S_{Miss} - S \mid U) &= E(S_{Miss} \mid U) - E(S \mid U) && (3.20)\\
&= \sum_{i=1}^{I} P(Y_i = 1 \mid U) \cdot P(D_i = 1 \mid U) - \sum_{i=1}^{I} P(Y_i = 1 \mid U) && (3.21)\\
&= \sum_{i=1}^{I} P(Y_i = 1 \mid U) \cdot P(D_i = 1 \mid U) - P(Y_i = 1 \mid U) && (3.22)\\
&= \sum_{i=1}^{I} [P(D_i = 1 \mid U) - 1] \cdot P(Y_i = 1 \mid U) && (3.23)\\
&= -\sum_{i=1}^{I} P(D_i = 0 \mid U) \cdot P(Y_i = 1 \mid U) && (3.24)
\end{aligned}
$$

Evidently, the expected sum score of any person u of U will be biased if $P(D_i = 0 \mid U) > 0$ for any item i. Equivalent to 3.19 this proves that the sum score is only expected to be unbiased when no missing data exist. Of course, so far we assumed implicitly that

each person has a positive probability $P(Y_i = 1 \mid U)$. In one-, two-, and three-parameter logistic IRT models this is equal to the assumption that each u of U has a value $\xi > -\infty$. In fact, from Equations 3.19 and 3.24 follows that in presence of any missing data mechanism the sum score is only unbiased if $P(Y_i = 1 \mid U) = 0$. This is at least the case if $Y_i \perp D_I \mid U$ holds true. However, if conditional stochastic dependence $Y_i \not\perp D_I \mid U$ exists, then the derivations above are not correct. This case can be be studied by rewriting the regression $E(S_{Miss} \mid U) = \sum_{i=1}^{I} P(Y_i = 1, D_i = 1 \mid U)$. Inserting this term into Equation 3.22 yields

$$
\begin{aligned}
E(S_{Miss} - S \mid U) &= \sum_{i=1}^{I} P(Y_i = 1, D_i = 1 \mid U) - P(Y_i = 1 \mid U) && (3.25)\\
&= \sum_{i=1}^{I} P(D_i = 1 \mid Y_i = 1, U) \cdot P(Y_i = 1 \mid U) - P(Y_i = 1 \mid U) && (3.26)\\
&= \sum_{i=1}^{I} [P(D_i = 1 \mid Y_i = 1, U) - 1] \cdot P(Y_i = 1 \mid U) && (3.27)\\
&= -\sum_{i=1}^{I} P(D_i = 0 \mid Y_i = 1, U) \cdot P(Y_i = 1 \mid U) && (3.28)\\
&= -\sum_{i=1}^{I} P(D_i = 0, Y_i = 1 \mid U) && (3.29)\\
&= -\sum_{i=1}^{I} P(Y_i = 1 \mid D_i = 0, U) \cdot P(D_i = 0 \mid U). && (3.30)
\end{aligned}
$$

Hence, in presence of any missing data mechanism with respect to at least one item i the sum score is only unbiased if the probability to solve a missing item given U is zero. This is implausible in almost all real applications and would have awkward implications. If a latent trait model applied with $\xi = f(U)$ exists, Equation 3.30 implies $P(Y_i = 1 \mid D_i = 0, \xi) = 0$ for all missing items irrespective of their item difficulty and the value of the latent variable of the person. This, in turn, implies $Y_i \perp \xi \mid D_i = 0$. This is a very strong form of differential item functioning since the model of Y_i depends on D_i. If $D_i = 1$ the latent trait model with $P(Y_i = 1 \mid \xi)$ holds. However, this model cannot be valid if $D_i = 0$ unless $\xi = -\infty$. In the latter case, however, all other observed item responses needs to be zero given the model is correct. In other words, assuming that Equation 3.30 holds means that any latent trait model is assumed only to be valid to observed responses. This

implication is typically ignored. That is worrisome since scoring missing responses as wrong is still commonly used in many assessments, which utilize IRT models. This so-called Incorrect-Answer-Substitution (IAS) will be considered in more detail in Section 4.3.1 using the derivations from this section.

Figure 3.2 shows the expected sum scores $E(S \mid U)$ and $E(S_{Miss} \mid U)$ of Data Example A. The correlation $Cor[E(S \mid U), E(S_{Miss} \mid U)] = 0.964$. Insofar, the high correlation in the

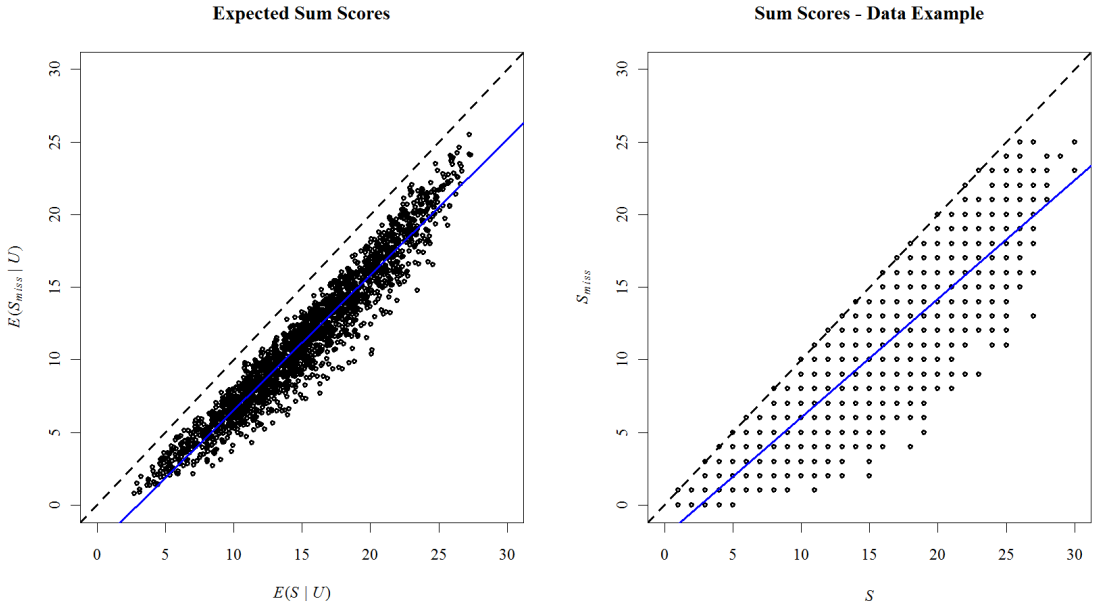

Figure 3.2: Comparison between the expected sum scores $E(S \mid U)$ and $E(S_{Miss} \mid U)$ (left) and the sum scores S and S_{Miss} (right) in Data Example A. The grey dotted line is the bisectric and the blue line is the regression line.

data example seems to suggest that the rank order is not affected. However, this is specific for the conditions used to simulate this particular data example. The high correlation is driven by the strong covariance between the ξ and θ. The lower $Cor(\xi, \theta)$ is, the higher the probability is that even highly proficient persons show considerable proportions of missing data. And the bias is expected to increase with increasing values of ξ since the omitted items are more likely to be answered correctly due to higher probabilities $P(Y_i = 1 \mid \xi)$. Non-responses in low proficient persons are less influential with respect to the bias of the sum score. Their probabilities $P(Y_i = 1 \mid \xi)$ are comparably low. From Equation 3.24 follows that the bias is generally small given $P(Y_i = 1 \mid \xi)$ is small. For the purpose of illustration, two additional data examples with the same 30 items and the same sample size were generated to show the effect of lower correlations between ξ and θ.

Two conditions were simulated $Cor(\xi, \theta) = 0.5$ and $Cor(\xi, \theta) = 0.2$. Figure 3.3 illustrate the effects graphically. The correlations $Cor[E(S \mid U), E(S_{Miss} \mid U)]$ of the expected sum scores are 0.902 given $Cor(\xi, \theta) = 0.5$ and 0.792 if $Cor(\xi, \theta) = 0.2$. The correlations were even lower for the realized sum scores in both simulations ($r(S, S_{Miss}) = 0.815$ given $Cor(\xi, \theta) = 0.5$; $r(S, S_{Miss}) = 0.723$ given $Cor(\xi, \theta) = 0.2$). Consequently, the correlation $Cor(S_{Miss}, S)$ decreases as well with decreasing values $Cor(\xi, \theta)$. This implies, in turn, that the reliability decreases too. However, even if the correlation $Cor(S_{Miss}, S)$

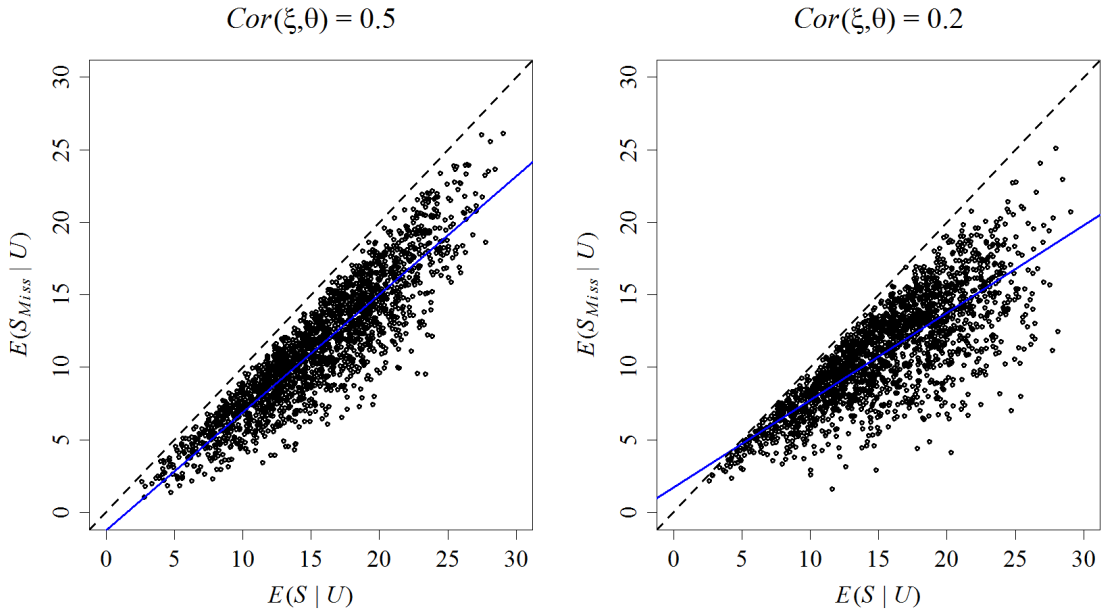

Figure 3.3: Comparison between expected sum scores $E(S \mid U)$ and $E(S_{Miss} \mid U)$ given $Cor(\xi, \theta) = 0.5$ (left) and $Cor(\xi, \theta) = 0.2$ (right). The grey dotted line is the bisectric. The blue line is the regression line.

is very high and the reliability and the rank order are hardly affected, the expected value of the sum scores $E(S_{Miss})$ can be considerably shifted. Since CTT is a norm-referenced assessment, this threatens the interpretation of test scores. For example, assume that there were two assessments of the same population: a low- and high-stakes assessment. As in real testings, the rates of missing responses were much larger in low-stakes than in high-stakes assessments. Data of the low-stakes assessment were used for standardization. If the test scores S_{Miss} or monotone functions $f(S_{Miss})$ of the high-stakes assessment were interpreted with respect to these test norms, the sample of the high-stakes assessment would seem to be more proficient because of lower rates of item nonresponses. The standardization group was tested under typical low-stakes conditions. If this is related to

higher proportion of missing data, the test norms are not meaningful in the high stakes assessment.

To summarize this section, simply to use the number right score as test score means to consider different variables depending on the presence of missing data. The sum score S in absence of non-responses and S_{Miss} in presence of missing data are different random variables with different distributions. Whereas S depends only on the items Y_i, S_{Miss} is a sum of the I product variables $(Y_i \cdot D_i)$. It was shown that the sum score is always negatively biased due to the implicit recoding of non-responses to $Y_i = 0$. In achievement tests this means that missing responses are treated as observed incorrect answers. From the statistical point of view this ignores the positive probability of a correct response given the latent ability even for omitted or not reached items. This was shown analytically considering the conditional expected bias $E(S_{Miss} - S \mid U)$ given the person variable. Assuming a latent response propensity θ the correlation $Cor(S_{Miss}, S)$ decreases with lower correlations $Cor(\xi, \theta)$ which results in lower reliabilities of S_{Miss}. Finally, the test norms become meaningless if the missing data mechanism and the distribution of $\mathbf{D}$ differs between the standardization group and the sample of interest, even if both are representative samples with respect to the latent variable that is intended to be measured. It was noted that the treatment of missing items as incorrect responses is implicit using the sum score but is explicit when incorrect answer substitution is applied. This is still widely used in applications of latent trait models and will be examined in more detail in Section 4.3.1.

3.1.2 Proportion correct

Using the sum score when missing data are present is equivalent to recoding non-responses to zero. It was shown that in practically all real situations a negative bias will result unless very strong and implausible assumptions hold true. Due to plausibility considerations, often the proportions correct score P^+ is preferred to the sum score, because the number of missing responses is taken into account. P^+ is defined as

$$P^+ = \frac{\sum_{i=1}^{I} Y_{i,D_i=1}}{\sum_{i=1}^{I} D_i}, \tag{3.31}$$

given that at least one item is responded to $(\sum_{i=1}^{I} D_i \geq 1)$. Suppose that two test takers u_1 and u_2 answered 10 items correctly but u_1 completed 30 items whereas u_2 answered 50 items. A comparison of the achievement between the two examinees based on the sum score would suggest equal performance on the test. Taking into account that u_1 answered

only 30 items the proportion of correctly answered items is $P^+(u_1) = 1/3$ compared to $P^+(u_2) = 1/5$ of person u_2. Obviously, the conclusion would be different depending on the test score S or P^+. At first sight, it seems plausible to prefer P^+, because P^+ is a individually standardized sum score. This can directly be seen in Equation 3.31. The nominator is simply the sum score S_{Miss}, that is scaled by the number of completed items $\sum_{i=1}^{I} D_i$ in the denominator of Equation 3.31. Therefore, P^+ accounts for missingness and does not implicitly convert missing values into $Y_i = 0$. The question is whether the standardization by the number of completed items is sufficient to accomplish comparability between test takers.

In order to answer this question, we could proceed similarly as in the case of the sum score. That is, the expected proportion correct $E(P^+ \mid U)$ can be considered in absence and in presence of a nonresponse mechanism. If no missing data mechanism exists, then $E(P^+ \mid U)$ is simply $I^{-1} \cdot E(S \mid U)$ since all response indicators are $D_i = 1$. However, under any missing data mechanism the number of answered items $\sum_{i=1}^{I} D_i$ is also a random variable. Generally the regression $E(P^+ \mid U)$ can be written as the conditional expectation of Equation 3.31 given U:

$$E(P^+ \mid U) = E\left(\frac{1}{\sum_{i=1}^{I} D_i} \cdot \sum_{i=1}^{I} Y_{i,D_i=1} \,\middle|\, U \right) \tag{3.32}$$

Let $W = (\sum_{i=1}^{I} D_i)^{-1}$ be the number of answered items. The nominator of Equation 3.31 is equal to S_{Miss} (cf. Equation 3.17). Therefore, we can rewrite Equation 3.32 as

$$\begin{aligned} E(P^+ \mid U) &= E(W \cdot S_{Miss} \mid U) && (3.33) \\ &= E(W \mid U) \cdot E(S_{Miss} \mid U) + Cov(W, S_{Miss} \mid U) && (3.34) \end{aligned}$$

Let ε_W and $\varepsilon_{S_{Miss}}$ be the residuals of $E(W \mid U)$ and $E(S_{Miss} \mid U)$ respectively. The conditional covariance $Cov(W, S_{Miss} \mid U)$ equals the regression $E([W - E(W \mid U)][S_{Miss} - E(S_{Miss} \mid U)] \mid U)$. Because $E(\varepsilon_W) = E(\varepsilon_{S_{Miss}}) = 0$, this is the conditional covariance $Cov(\varepsilon_W, \varepsilon_{S_{Miss}} \mid U)$ of the residuals. Assuming $Cov(\varepsilon_W \cdot \varepsilon_{S_{Miss}} \mid U) = 0$, it follows:

$$E(P^+ \mid U) = E(W \mid U) \cdot E(S_{Miss} \mid U). \tag{3.35}$$

The first regression is the expected inverse number of answered items given the person variable U. Unfortunately, from the Jensen's inequality follows $E[f(\sum_{i=1}^{I} D_i)] >$

$f[E(\sum_{i=1}^{I} D_i)]$ (Heijmans, 1999; Koop, 1972). Hence, Equation 3.35 can only be simplified to:

$$E(P^+ \mid U) = E\left[\left(\sum_{i=1}^{I} D_i\right)^{-1} \middle| U\right] \cdot \sum_{i=1}^{I} P(Y_i = 1 \mid U) \cdot P(D_i = 1 \mid U) \tag{3.36}$$

Nevertheless, this Equation is insightful with respect to the expected bias of the proportion correct score. The regression $E(S_{Miss} \mid U)$ is a weighted sum. The true scores [3] are weighted by the item response propensities $P(D_i = 1 \mid U)$. If easier items are more likely to be answered, the values of the regression $E(S_{Miss} \mid U)$ will be higher than in situations when difficult items are preferred to be answered. However, the expectation $E(W \mid U)$ of the reciprocal of the number of completed items given U does not account for differences in characteristics of items that are more or less likely answered. From this point of view, the proportion correct score can be positively or negatively biased. If easier items are more likely to be completed by test takers while difficult items are preferentially omitted, the proportion correct score is expected to be positively biased. In contrast, if there is a tendency to skip easier items while preferring to answer difficult items, the P^+ is most likely negatively biased. If a person with a given ability chooses only easy items, the expected proportion correct will be higher than when completing a selection of only difficult items.

In previous studies is has become evident that in educational low stakes assessments preferentially more difficult items are omitted (Culbertson, 2011, April; Rose et al., 2010). This might reflect psychological evaluative processes of test takers while completing the test. At least in achievement tests, it seems that examinees judge the difficulty of the items. More likely such items are completed that are expected to be answered correctly. As a consequence, more difficult items are more likely skipped. In order to study the effects of systematic selection of items depending on their difficulties, the mean test difficulty T_β can be considered. T_β is the mean of the item difficulties of those items answered by a test taker. That is

$$T_\beta = \frac{\sum_{i=1}^{I} \beta_i \cdot D_i}{\sum_{i=1}^{I} D_i}. \tag{3.37}$$

T_β can be calculated for each test taker. If no missing data mechanism exist T_β is a constant $T_\beta = I^{-1} \sum_{i=1}^{I} \beta_i$. However, if a nonresponse mechanism exists, T_β is the mean item difficulty of only the completed items and is a measure of the average difficulty of the

[3] Since $P(Y_i = 1 \mid U) = \tau_i$.

test with item non-responses. If a test taker omitted only difficult items, T_β will be low. Omissions of only easy items result in a high value of T_β. The average test difficulty can and will most likely vary across the persons depending on the missing pattern. However, the comparability of test scores is in doubt if each test taker composes his or her own test consisting of different items. T_β is of diagnostic value. It can be utilized to study examinees choice behaviour of items with respect to its difficulty. If the item response propensities $P(D_i = 1 \mid U = u)$ are known for each u of U, the weighted mean $T_\beta^{(w)}$ can be computed and is given by

$$T_\beta^{(w)} = \frac{\sum_{i=1}^{I} \beta_i \cdot P(D_i = 1 \mid U)}{\sum_{i=1}^{I} P(D_i = 1 \mid U)}. \tag{3.38}$$

Whereas T_β is a function $f(\boldsymbol{\beta}, \boldsymbol{D})$, $T_\beta^{(w)}$ is a function $f(\boldsymbol{\beta}, U)$ of the item difficulties and the person variable U that can be interpreted as an approximation of the expected mean test difficulty of a person[4]. As already noted, it is expected that T_β and $T_\beta^{(w)}$ vary across the persons. If easier items are generally preferred by test takers and $Cor(\xi, \theta) \neq 0$, then a systematic relationship between the latent ability ξ and T_β as well as between ξ and $T_\beta^{(w)}$ is implied. Hence, examinees prefer to skip too-difficult items relative to their ability. The test takers compose their own test with items they expect to respond to correctly. Figure 3.4 shows T_β and $T_\beta^{(w)}$ given the latent response propensity and the latent ability. In Data Example A a latent response propensity θ as a function $f(U)$ was used to determine item response propensities. Therefore, $P(D_i = 1 \mid U)$ in Equation 3.38 was replaced by $P(D_i = 1 \mid \theta)$ implying that $T_\beta^{(w)} = f(\boldsymbol{\beta}, \theta)$ (red line in Figure 3.4). As expected the mean test difficulty decreases with lowering values of the latent response propensity. Due to the high correlation $Cor(\xi, \theta) = 0.8$, the expected mean test difficulties $T_\beta^{(w)}$ and T_β are also strongly correlated with ξ ($r(T_\beta^{(w)}, \xi) = 0.797$ and $r(T_\beta, \xi) = 0.548$).

However, the weighted mean test difficulty $T_\beta^{(w)}$ is not necessarily a strictly monotonically increasing function of θ as in Data Example A. Equation 3.38 shows that $T_\beta^{(w)}$ is determined by item difficulties β_i and item response propensities $P(D_i = 1 \mid U)$. Considering the case where the item response propensities are a parametric function of a latent response propensity θ with $P(D_i = 1 \mid U) = P(D_i = 1 \mid \theta)$, $T_\beta^{(w)}$ is a function $f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta)$. In this case $\boldsymbol{\gamma}$ denotes the vector of parameters of the model of $\boldsymbol{D}$. In Data Example A this is the vector $\boldsymbol{\gamma} = \gamma_1, \ldots, \gamma_I$ of thresholds (see Equations 3.2). In this case $T_\beta^{(w)}$ depends on

[4]Strictly speaking the expected mean test difficulty is $E(T_\beta \mid U) = E[(\sum_{i=1}^{I} D_i)^{-1} \sum_{i=1}^{I} \beta_i \cdot D_i \mid U)$ that is again the expectation of a ratio and is not exactly equal to the weighted mean $T_\beta^{(w)}$ (Heijmans, 1999; Koop, 1972).

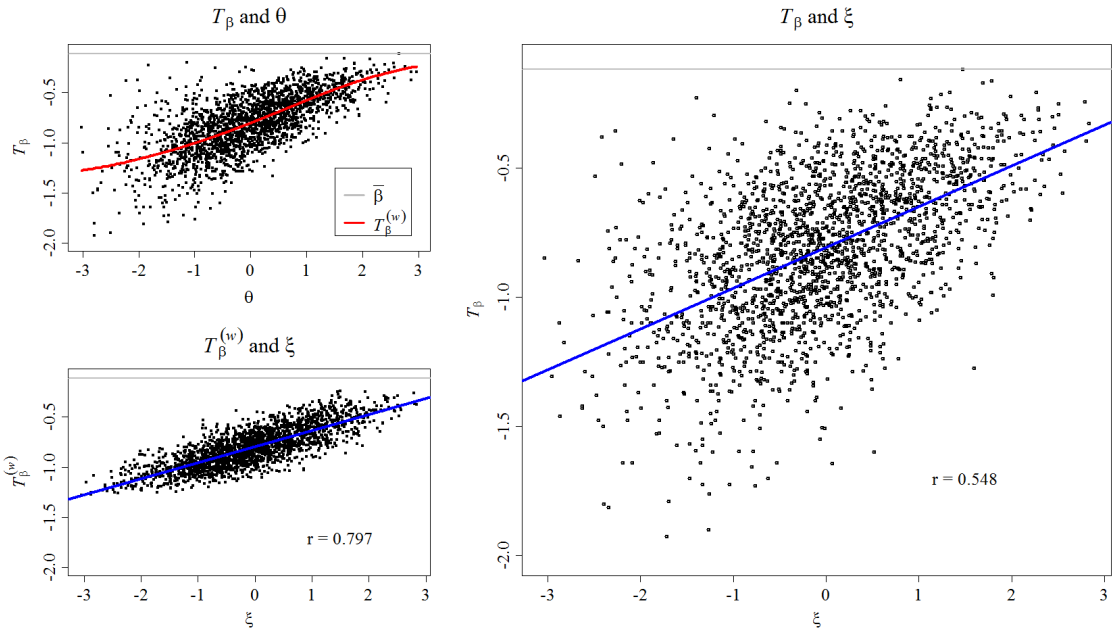

Figure 3.4: Relationship between individual mean test difficulties (T_β and $T_\beta^{(w)}$) and the latent variables ξ and θ (Data Example A). The grey line represents the mean $\bar{\beta}$. The blue line is the regression line.

the three factors: (a) the latent response propensity, (b) the parameters of the regression $P(D_i = 1 \mid \theta)$, and (c) the item difficulties β_i.

To study the influence of these factors on $T_\beta^{(w)}$, different cases can be considered theoretically. First, it is assumed that all parameters γ_i are equal for all D_i implying $P(D_i = 1) = P(D_j = 1)$ (for all i and j in $1, \ldots, I$). Nevertheless $P(D_i = 1 \mid \theta)$ may vary across the persons due to interindividual differences in the latent response propensity. However, given a particular person u with $\theta(u)$, the item response propensities are equal across the items. In this case the index i can be omitted: $P(D_i = 1 \mid \theta) = P(D = 1 \mid \theta)$ (for all $i = 1, \ldots, I$). Equation 3.38 of the weighted mean test difficulty can be written as

$$T_\beta^{(w)} = \frac{\sum_{i=1}^{I} \beta_i \cdot P(D_i = 1 \mid \theta = \theta)}{\sum_{i=1}^{I} P(D_i = 1 \mid \theta = \theta)} \tag{3.39}$$

$$= \frac{\sum_{i=1}^{I} \beta_i \cdot P(D = 1 \mid \theta = \theta)}{\sum_{i=1}^{I} P(D = 1 \mid \theta = \theta)} \tag{3.40}$$

$$= \frac{P(D = 1 \mid \theta = \theta) \cdot \sum_{i=1}^{I} \beta_i}{I \cdot P(D = 1 \mid \theta = \theta)} \tag{3.41}$$

$$= \frac{\sum_{i=1}^{I} \beta_i}{I}. \tag{3.42}$$

Hence, if the parameters are equal for all items, then the weighted mean test difficulty $T_\beta^{(w)}$ is constant and equal to the unconditional mean $\bar{\beta}$ of item difficulties. If $\boldsymbol{D} \perp \xi \,|\, \theta$ this additionally implies that the weighted mean test difficulty is always $\bar{\beta}$ regardless the value of the latent ability ξ of the test takers. Hence, if a latent response propensity exist the equality of parameters γ_i across the response indicators suggests that persons do not tend to omit items in a way such that the average difficulty depends on the latent ability. However, in realized data T_β can vary depending on the realized missing data pattern $\boldsymbol{D} = \boldsymbol{d}$.

A second case where $T_\beta^{(w)}$ is constant across persons is trivial. If the item difficulties β_i are equal for all items i in $1, \ldots, I$, then the index i can be omitted from difficulty parameters β. Hence

$$
\begin{aligned}
T_\beta^{(w)} &= \frac{\sum_{i=1}^{I} \beta \cdot P(D_i = 1 \,|\, \theta)}{\sum_{i=1}^{I} P(D_i = 1 \,|\, \theta)} && (3.43)\\
&= \frac{\beta \cdot \sum_{i=1}^{I} P(D_i = 1 \,|\, \theta)}{\sum_{i=1}^{I} P(D_i = 1 \,|\, \theta)} && (3.44)\\
&= \beta. && (3.45)
\end{aligned}
$$

Thus, if all i items have the same difficulty, then $T_\beta^{(w)} = T_\beta = \beta$.

The theoretical considerations highlight that the stochastic relation between T_β and the latent variable ξ is mainly driven by the correlation $Cor(\theta, \xi)$ and the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ of a parametric model for $(\boldsymbol{Y}, \boldsymbol{D})$. To illustrate these findings, additional data sets were simulated. Figure 3.5 shows the results for different correlations $Cor(\xi, \theta)$ and varying magnitudes of the relation between the parameters γ_i and β_i. To express this relationship between the parameter vectors by a single value, the sample correlation coefficient $r(\beta, \gamma)$ was used. It is important to note that the correlation is defined with respect to two random variables. The parameter vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are typically not considered to be vectors of identical and independently distributed random variables. However, the sample correlation coefficient $r(\beta, \gamma)$ is computed in the same way as the sample estimate of the correlation and is given by

$$
r(\beta, \gamma) = \frac{\sum_{i=1}^{I} (\beta_i - \bar{\beta}) \cdot (\gamma_i - \bar{\gamma})}{\sqrt{\sum_{i=1}^{I} (\beta_i - \bar{\beta})^2} \cdot \sqrt{\sum_{i=1}^{I} (\gamma_i - \bar{\gamma})^2}}. \qquad (3.46)
$$

$r(\beta, \gamma)$ is useful here to express the relationship between difficulties of items and their overall chance to be answered or not, which is expressed by γ_i. In Equation 3.46 $\bar{\beta}$ and $\bar{\gamma}$

are the means of the respective parameters β_i and γ_i. The nine simulated data examples were simulated with the same parameters β_i as shown in Table 3.1. The parameters γ_i are different but correlated with β_i. The values 0, 0.5, and 0.8 were chosen for $Cor(\xi, \theta)$, and 0.08, 0.46, and 0.95 were chosen for $r(\beta, \gamma)$. Hence,the easier the items are, the higher the unconditional probabilities of an item response are. The overall proportion of missing data ranged between $47 - 49\%$ similar to Data Example A. The direction of the correlation $r(\beta, \gamma)$ determines also the direction of the correlation $Cor(T_\beta, \theta)$. The correlation was always $r(\beta, \gamma) > 0$. The direction of the correlation $Cor(T_\beta, \xi)$ depends on both $Cor(T_\beta, \theta)$ and $r(\beta, \gamma)$. If $Cor(\xi, \theta) > 0$ and $Cor(T_\beta, \theta) > 0$ or if both are negative, then $Cor(T_\beta, \xi)$ will be positive. If the correlations $Cor(T_\beta, \theta)$ or $r(\beta, \gamma)$ have oppositional signs, then $Cor(T_\beta, \xi)$ will be negative. Figure 3.5 allows to conclude that the correlation

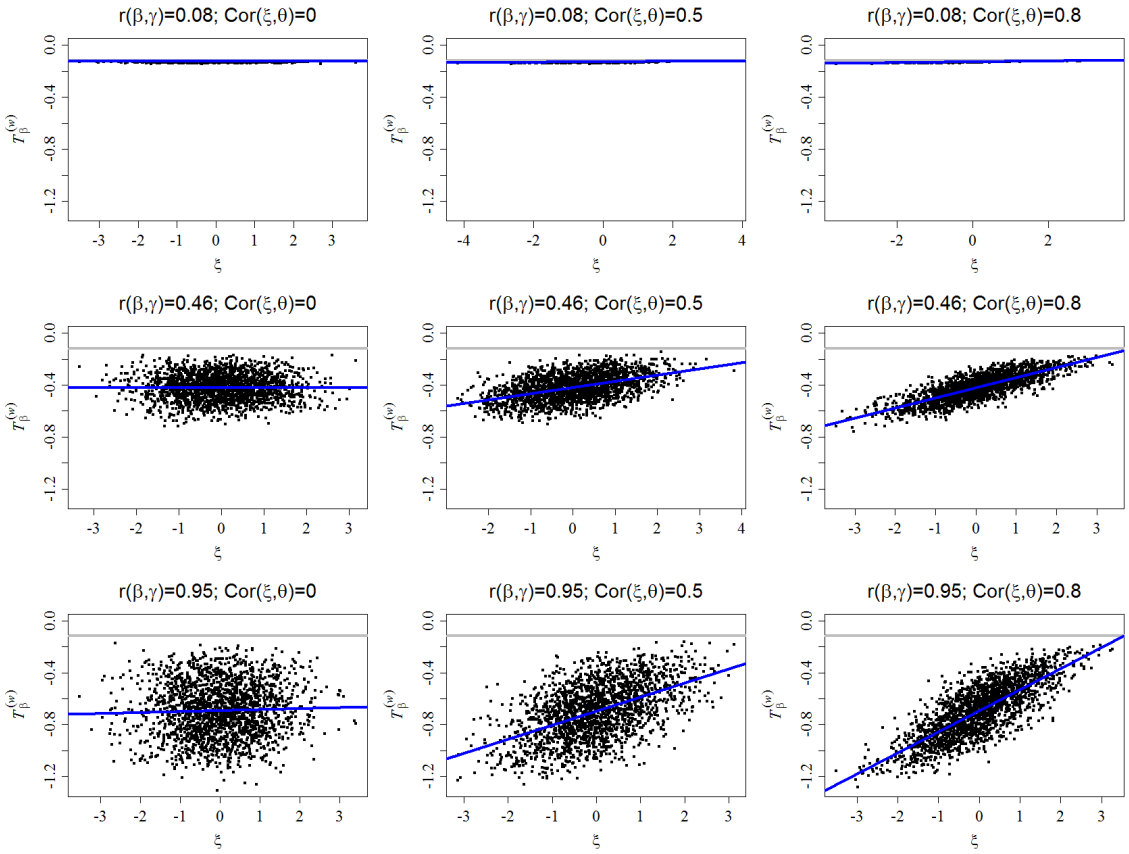

Figure 3.5: Nine simulated data sets with different values for $Cor(\xi, \theta)$ and $r(\beta, \gamma)$. The blue line represents the linear regression $E(T_\beta^{(w)} | \xi)$. The gray line indicates the mean $\bar{\beta}$ of the item difficulties.

$r(\beta, \gamma)$ is strongly related to the average drop of the mean test difficulty due to item selec-

tion. Whereas, the correlation $Cor(\xi, \theta)$ drives the relationship between the latent ability ξ and missingness, the relationship between β and γ determines how systematic item nonresponses are with respect to items? difficulties. If both $Cor(\xi, \theta)$ and $r(\beta, \gamma)$ are high, the mean test difficulty is strongly related to ξ. What are the implication of these results with respect to the use of P^+ in presence of missing data?

The findings demonstrated that the omission of items means that test takers compose their own test. The P^+ score does not account for this selection. This is problematic if items are systematically skipped due to their characteristics such as the item difficulty. For example, if preferably difficult items are omitted the average test difficulty T_β decreases. If contemporarily the response propensity of test takers is correlated with the ability, then the mean test difficulty is also positively correlated with the latent ability ξ. Hence, the lower the proficiency levels of a person is, the higher the probability of responding only to easy items while skipping difficult items. Since the P^+ score only accounts for the number of omitted items but not which items are missing, the bias of P^+ can be positive or negative. If $T_\beta > \bar{\beta}$, then the bias of P^+ is expected to be negative. If $T_\beta < \bar{\beta}$, then the bias of P^+ is expected to be positive. Data Example A was generated so that $T_\beta \leq \bar{\beta}$ for all test takers. This is in line with most empirical findings. Difficult items are most likely to be missing and the tendency to produce item nonresponses and the persons' proficiency is positively correlated. In this case the bias of P^+ should be positive especially for persons with lower ability levels. The bias of the proportion correct score is given by $P^+ - S/I$. Note that S/I is the proportion correct without missing data which is typically not available in real applications with missing data. Additionally, the expected bias $E(P^+ \mid U) - I^{-1} \cdot E(S \mid U)$ can be considered. As Equation 3.36 shows, the conditional expectation $E(P^+ \mid U)$ involves the regression $E[(\sum_{i=1}^{I} D_i)^{-1} \mid U]$ whose values are difficult to obtain. For Data Example A, the values $E[(\sum_{i=1}^{I} D_i)^{-1} \mid U = u]$ were approximated by simulating 1000 data sets with the true person parameters of ξ and θ and the true item parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. For each test taker 1000 simulated missing patterns resulted. The means of the inverse sums of completed items were used as estimates $\hat{E}[(\sum_{i=1}^{I} D_i)^{-1} \mid U = u]$. In the next step, these values were inserted in Equation 3.36 to obtain approximations of $E(P^+ \mid U = u)$. Finally, the expected bias of P^+ was computed. Figure 3.6 shows the expected and the observed bias of the proportion correct score as realized in Data Example A, in relation to θ ad ξ. The bias increases with a lower willingness or tendency of the examinees to respond to test items. As expected, only to complete easy items is beneficial for most test takers. In other words, the omission of difficult items is rewarded when the proportion correct is used as test scores. Note that persons with very low values of ξ will not profit from

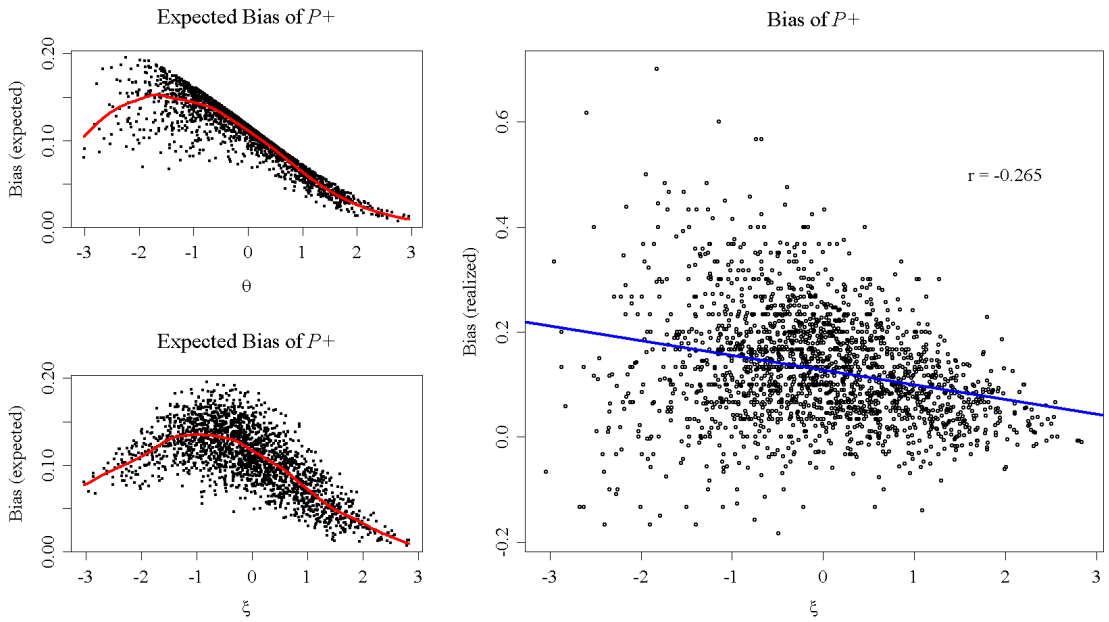

Figure 3.6: Expected and observed bias of the proportion correct score P^+ given θ and ξ (Data Example A). The red line represents a smoothing spline regression. The blue line denotes a linear regression.

selecting easy items, since the probability of a correct responses are quite low even for these items. Similarly, highly proficient persons also show very little bias because even the difficult items, which most likely would be skipped, would be answered correctly by most of these persons. Finally, in the left graph of Figure 3.6, the bias of the observed data are shown in relation to the latent ability ξ.

It is important to note that some these results are specific for the investigated conditions used for the simulation of Data Example A. However, in conjunction with the theoretical derivations from above, the results allow for some general conclusions. Compared to the sum score, P^+ accounts for missing data by considering only completed items. However, this is not sufficient if test takers create their own test by selecting items due to certain item characteristics. If completed and missing items differ systematically, then the comparability of the proportion correct scores across examinees potentially get lost. For example, if items are picked out by test takers due to item difficulties, then the proportion correct score P^+ will be biased. Theoretically, the bias can be positive or negative depending on whether easy or difficult items are preferentially omitted or not reached. Here the mean test difficulty T_β has been introduced to quantify item selection due to item difficulties. In applications T_β can easily be estimated using the item parameter estimates $\hat{\beta}_i$. This might

be of diagnostic value in order to study systematics in selection of items. Data Example A was simulated in accordance with empirical findings that report that preferentially difficult items are more likely skipped. In this case, P^+ tends to reward the omission of items since P^+ is on average positively biased. This is all the more true, the stronger the relation between the difficulties β_i and $P(D_i = 0)$ is[5]. Since the proportion correct score P^+ does not account for systematic differences in observed and missing items, the use of P^+ seems questionable in most applications regardless of whether the missing data mechanism is MCAR, MAR, or non-ignorable. There are only a few less realistic situations where the use of P^+ is unproblematic in presence of missing data. Only if all items have the same item difficulties, then the proportion correct scores are comparable across persons with different missing pattern. Hence, the use of P^+ as the test score is not recommended under any missing data mechanism.

3.1.3 IRT Based Test Scores: MLE, WLE, and EAP

Different estimation methods have been developed in order to obtain estimates for persons' individual trait levels as well as item parameters. The joint maximum likelihood estimation (JML; e. g. Baker & Kim, 2004) has been developed first and has its roots in the fundamental work of Birnbaum (Birnbaum, 1968). JML can be used for one-, two-, and three-parametric IRT models. Unfortunately, the JML estimation suffers from inconsistent parameter estimates since the number of estimates increases with the number of observations (e g. Little & Rubin, 1984; Baker & Kim, 2004). The problem of inconsistency can be circumvented using the conditional maximum likelihood (CML) method. CML is based on the property that the sum score is a sufficient statistic with respect to the latent person variable $\boldsymbol{\xi}$ in one-parameter models of the Rasch family[6]. Unfortunately, CML is not applicable for two- and three-parameter models. With the marginal maximum likelihood (MML) estimation method an alternative ML estimator has been developed for one-, two-, and three-parameter models (Bock & Lieberman, 1970; Bock & Aitkin, 1981; Baker & Kim, 2004). The problem of inconsistency is solved by assuming a distribution of the latent variable $\boldsymbol{\xi}$ that can be described parametrically. Instead of estimating all person parameters, only the parameters of the distribution $g(\boldsymbol{\xi})$ need to be estimated jointly with the item parameters. The number of estimands is then independent of the

[5] In Data Example A $P(D_i = 0)$ depends on and the parameters γ_i relative to the distribution of the latent response propensity. Therefore, the relationship between β_i and $P(D_i = 0)$ is reflected by $r(\gamma, \beta)$

[6] The Rasch family subsumes models for dichotomous or polytomous items where the item discrimination parameter equals one and the lower asymptote (*„pseudo-guessing parameter“*) is zero.

sample size. Usually, the (multivariate) normal distribution is assumed, which is sufficiently specified by the vector of expected values $E(\boldsymbol{\xi})$ and variance-covariance matrix $\Sigma(\boldsymbol{\xi})$. The advantage of consistent item parameter estimation under MML is offset by additional estimation stages required to obtain individual person parameter estimates. These are estimated in a subsequent estimation procedure taking the previously estimated item parameters as known. Different ability estimators have been developed. Here the consideration is confined to three estimators commonly used in educational and psychological testings: (a) the ML estimate, (b) Warm's weighted maximum likelihood (WML) estimate, and (c) the expected a posteriori (EAP) estimate. Due to the outlined shortcomings of inconsistent parameter estimates, JML estimation will be left out here. Hence, the bias of ML, WML, and EAP person parameter estimates is confined to the case where item parameters are estimated with MML ignoring missing data in a first step and subsequent estimation of person parameters in a second step based on incomplete response pattern $\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}}$. It should be noted that the generalizability of the results will be limited to MML estimation, since the bias of item and person parameter estimates due to missing data can be different depending on the estimation method, JML or MML (DeMars, 2002).

Since person parameter estimation under MML estimation is a two-step procedure that involves fixed item parameter estimates in the second step, unbiasedness of the person parameters rest upon unbiasedness of item parameter estimates. Biases that arise in earlier estimation stages are potentially transmitted to the subsequent person parameter estimation. As already mentioned, item and person parameter estimates are mutually dependent. That is, the ML estimators involves conditional response category probabilities $P(Y_i = y_i \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})$ to estimate both item and person parameters that are themselves functions of item parameters $\boldsymbol{\iota}$ and person parameters represented by $\boldsymbol{\xi}$. No closed-form expressions exist for estimation equations of item difficulties, item discriminations, and person parameters. Therefore, iterative estimation procedures such as the EM algorithm are required. As a consequence, in contrast to the sum score and the proportion correct score, analytical studies of the bias of item and person parameter estimates are quite limited. For that reason a simulation study was utilized to investigate potentially biased parameter estimation due to item nonresponses. The chosen conditions in the simulation study are described in the beginning of this chapter (see Chapter 3). Additionally, IRT parameter estimates of Data Example A will be presented for illustration.

Bias of IRT person parameter estimates due to item nonresponses Generally, all estimators under study were unweighted or weighted maximum likelihood or Bayesian

estimators. Rubin (1976) and Little and Rubin (2002) demonstrated in detail that ML and Bayesian estimators will be consistent and unbiased if the non-response mechanism is MCAR or MAR (ignorable missing data). Glas (2006) confirmed unbiased parameter estimation if the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y}$. De Mars (2002) stated that unbiased parameter estimation requires the inclusion of $\boldsymbol{Z}$ into a joint model of $(\boldsymbol{Y}, \boldsymbol{Z})$ if the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$. These issues will be discussed in detail in Section 4.5. For now it suffices to note that especially nonignorable item nonresponses are expected to result in biased parameter estimates. Therefore, the simulation study was confined to compare different conditions with nonignorable missing data and nonresponses that are MCAR. Covariates $\boldsymbol{Z}$ were not included. The degree of nonignorability was varied by different values of $Cor(\xi, \theta)$. If $Cor(\xi, \theta) = 0$, then the missing data mechanism w.r.t. $\boldsymbol{Y}$ was MCAR. The stronger the correlation $Cor(\xi, \theta)$ was, the stronger the implied stochastic dependency between $\boldsymbol{Y}$ and $\boldsymbol{D}$ was. In Section 2.3 it was scrutinized that $\boldsymbol{Y} \not\perp \boldsymbol{D}$ implies stochastic dependence between $\boldsymbol{D}$ and ξ suggesting that person parameter estimates are potentially biased. However, in contrast to the proportion correct score, IRT person parameter estimation includes information of the items that were completed. Hence, person parameter estimates are comparable across test takers even if different sets of items have been answered. Interindividual differences in T_β are not per se a problem and are sometimes even intended in branched and adaptive testing[7]. Therefore, neither the bias of the sum score nor the bias of the proportion correct score have direct implications with respect to the bias of IRT estimates regarding to individual values of ξ. Since person and item parameters have a common metric, the bias of the item parameters is expected to result in similarily biased person parameter estimates. Taken together, the following expectations can be formulated:

1. There is a systematic bias of ML, WML, and EAP estimates given the missing data with respect to $\boldsymbol{Y}$ is MNAR.

2. The pattern of biases of item and person parameters are expected to be similar.

The second expectation implies that the biases of item and person parameter estimates are correlated. Table 3.3 shows summary statistics of ML, WLM, and EAP estimates of Data Example A. The results were obtained using ConQuest 2.0 (Wu et al., 1998) for item and person parameter estimation.

[7]For example, in CAT T_β is expected to be correlated with ξ.

Table 3.3: Summary Information on ML-, WML-, and EAP Person Parameter Estimates Based on Complete and Incomplete Data (Data Example A).

Estimator	Mean	Variance	$r(\xi, \hat{\xi})$	$Rel(\hat{\xi})$	MSE	$r(bias, \xi)$
	Complete data					
ML	0.005	1.320	0.908	0.834	0.233	0.089
WML	0.002	1.241	0.910	0.826	0.214	0.027
EAP	0.000	0.859	0.912	0.833	0.169	-0.378
	Incomplete data (ignoring missingness)					
ML	-0.035	1.557	0.816	0.666	0.520	0.025
WML	-0.088	1.349	0.827	0.641	0.427	-0.061
EAP	-0.001	0.632	0.821	0.685	0.327	-0.608

Bias of ML person parameter estimates The ML estimation method is introduced in detail in Section 4.2. Here, it is sufficient to note that the ML person parameter estimate $\hat{\xi}_{ML}$ of a unidimensional latent variable ξ is that value $\xi \in \Omega_\xi$ that maximizes the conditional probability $P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{\xi})$. Ω_ξ is the parameter space. $\hat{\xi}_{ML}$ is estimated by maximizing the response pattern likelihood $\mathcal{L}(\boldsymbol{y_n}; \boldsymbol{\iota})$. The latter is proportional to the conditional probability $P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})$. The subscript n indicates that the response pattern likelihood is independently maximized for each test taker $n = 1, \ldots, N$. Under the assumption of local stochastic independence this is:

$$P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \xi; \boldsymbol{\iota}) \quad = \quad \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota}) \tag{3.47}$$

The pattern likelihood is a continuous differentiable function with respect to ξ. The value that maximizes $\mathcal{L}(\boldsymbol{y_n}; \boldsymbol{\iota})$ is the root of the first partial derivative of Equation 3.47 with respect to ξ. Due to theoretical and computational reasons, commonly the logarithm $\ln[\mathcal{L}(\boldsymbol{y_n}; \boldsymbol{\iota})] = \ell(\boldsymbol{y_n}; \boldsymbol{\iota})$ of the likelihood functions is used which is proportional to $\ln[P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \xi; \boldsymbol{\iota})]$ used for the estimation. For the case of dichotomous items Y_i, the first derivative of $\ell(\boldsymbol{y_n}; \boldsymbol{\iota})$ with respect to the person parameter is

$$\frac{\partial}{\partial \xi} \ln[P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \xi; \boldsymbol{\iota})] \quad = \quad \sum_{i=1}^{I} \alpha_i [y_{ni} - P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})]. \tag{3.48}$$

The estimation equation involves the I conditional probabilities $P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})$ that are themselves functions of the estimand ξ. Additionally, the item parameter indexed by $\boldsymbol{\iota}$ are

involved. In application the MML item parameter estimates are used instead of the true values $\boldsymbol{\iota}$. Hence, unbiased item parameter estimates are necessary for an accurate person parameter estimation. So far no missing data were considered. However, nonresponses reduce observed information for parameter estimation. In this case, only the likelihood $\mathcal{L}(\boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ of the observed response pattern is maximized. Formally, this can be expressed by the inclusion of response indicators including D_i in the Equations 3.47 and 3.53. The pattern likelihood of $\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}$ is proportional to

$$P(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \xi; \boldsymbol{\iota}) \quad = \quad \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi})^{D_i = d_i}. \tag{3.49}$$

Accordingly, the first derivative of $\ell(\boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ is proportional to

$$\frac{\partial}{\partial \xi} \ln[P(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \xi; \boldsymbol{\iota})] \quad = \quad \sum_{i=1}^{I} d_{ni} \alpha_i [y_i - P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota})]. \tag{3.50}$$

Hence, the response indicators work as selecting variables D_i. Only the weighted differences $\alpha_i[y_i - P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota})]$ of the responded items contribute to the estimation of the person parameter. The loss of information should be reflected by increased standard errors. Furthermore, ML estimates are potentially biased. Figure 3.7 shows the bias of the ML person parameter estimates in Data Example A given the latent variable ξ (left) and the number of nonresponses (right). Surprisingly, the bias of the ML estimates was uncorrelated with the latent ability ($r = 0.025$, $t = 1.109$, $df = 1998$, $p = 0.268$), although the number of item nonresponses was strongly correlated with the ξ ($r = -0.719$, $t = -46.274$, $df = 1998$, $p < 0.001$). Although difficult items are more likely to be skipped, on average persons do not profit from preferentially responding to easier items (see right graph of Figure 3.7). As expected, the variation of the ML estimates increased with the number of nonresponses but no systematic bias could be found. The correlation between the bias and the number of omitted items was not significant ($r = 0.025$,$t = 1.257$, $df = 1998$, $p = 0.209$). Additionally, Table 3.3 shows that the mean of the ML estimates is -0.035 ($t = -1.261$, $df = 1999$, $p = 0.207$). Since the latent variable ξ used to simulate the data was unit normally distributed, and the model was identified by fixing the scale of the latent variable to zero, the mean of $\hat{\xi}_{ML}$ is equal to the average bias of the ML estimator[8]. Hence, the average bias is not significantly different from zero. However, as the results of the simulation study revealed, this is not generally

[8] $E(\xi) = 0 \Rightarrow E(\hat{\xi}_{ML}) = E[\xi + Bias(\hat{\xi}_{ML})] = E(\xi) + E[Bias(\hat{\xi}_{ML})] = E[Bias(\hat{\xi}_{ML})]$

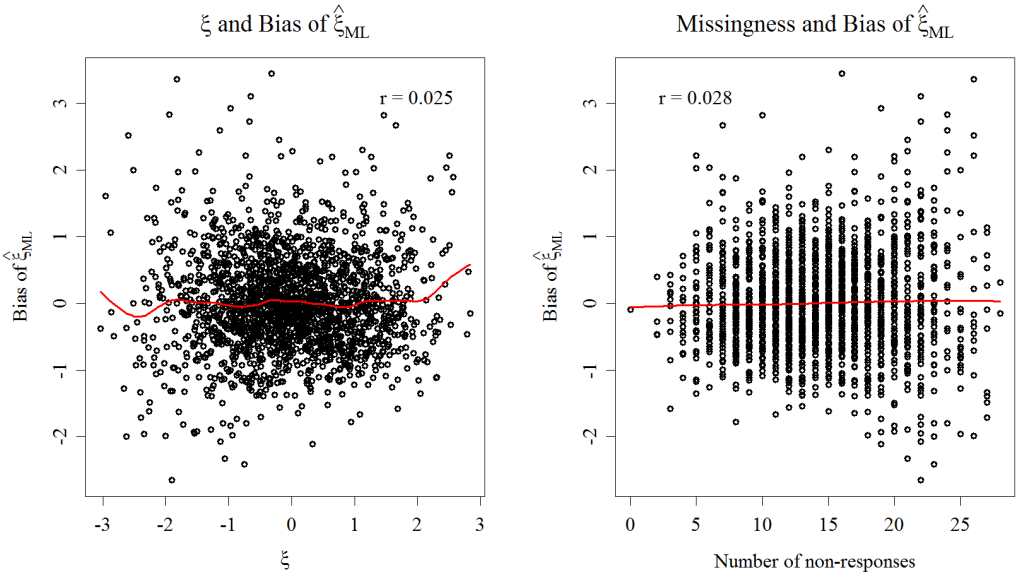

Figure 3.7: Relationship between the bias of the ML person parameter estimates of Data Example A and the latent variable ξ (left) and the number of non-responses (right). The red line represents a smoothing spline regression.

the case. Figure 3.8 shows a systematic negative bias of ML person parameter estimates. A largest bias was found if the correlation $Cor(\xi, \theta)$ was high and the overall proportion of missing data was large. The graph indicates an interaction between these two factors. Furthermore, there is a small positive bias if the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MCAR but the correlation $r(\beta, \gamma)$ increases. Nevertheless, the correlation $Cor(\xi, \theta)$ and the overall proportion of missing data seem to be the most influential factors determining the bias. This could be confirmed using a saturated regression model with the bias as the dependent variable and the factors shown in Table (see Table 3.2) as independent variables. Due to interaction terms the number of parameters is very large (720). The consideration and interpretation of single regression coefficients becomes challenging and may not facilitate the understanding of the importance of single factors with regard to the bias of ML estimates. Therefore, differences in R^2-values between regression models with and without particular factors are used to identify most important sources of the bias. To reduce the number of possible models, the seemingly most important two factors - the correlation $Cor(\xi, \theta)$ and the overall proportion of missing data - were given focus. Four saturated regression models were computed with the bias of the parameter estimates as dependent variable: (a) Model one (M_1) that contained all five factors that were systematically varied in the simulation study (see Table 3.2), (b) model two (M_2) without the factor $Cor(\xi, \theta)$,

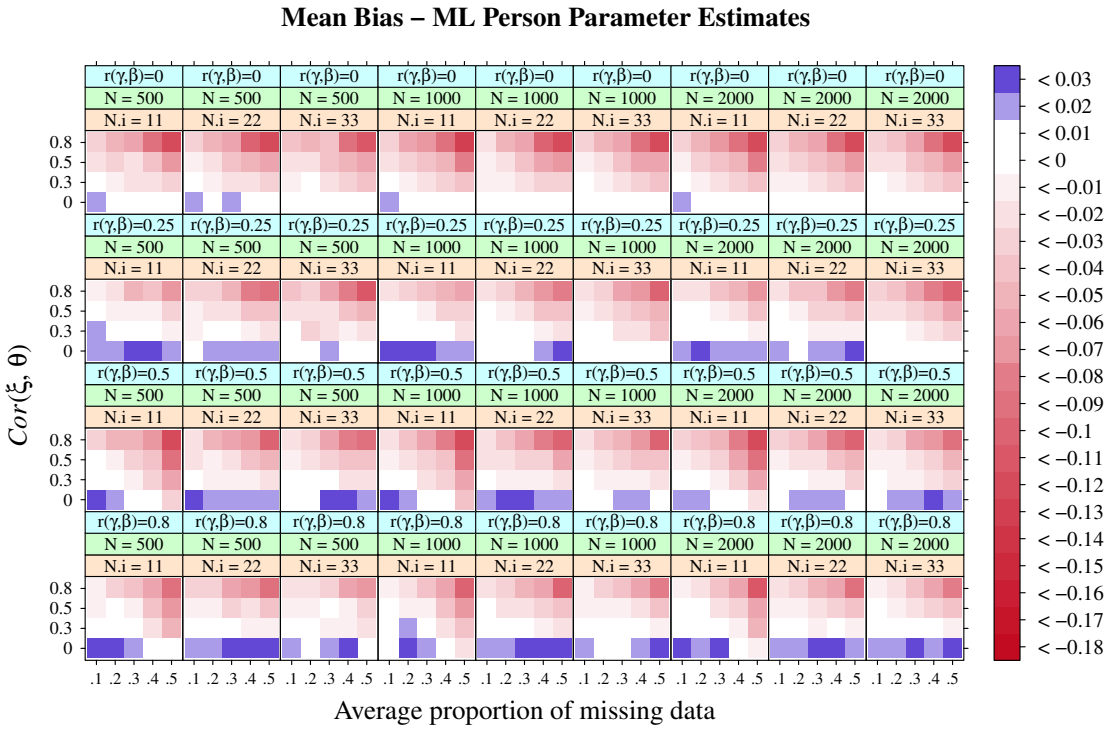

Figure 3.8: Mean Bias of the ML person parameter estimates using the 1PLM (simulation study).

(c) model three (M_3) where the overall proportion of missing data was not included as independent variable, and (d) model four (M_4) without both factors - $Cor(\xi, \theta)$ and the overall proportion of missing data. The results are summarized in Table 3.4. 38.9 % of the variation in the bias of ML person parameter estimates could be explained by all five factors in the simulation study. This proportion reduces to 10.7 % if $Cor(\xi, \theta)$ was not included as independent variable. 26.9 % explained variance was found in model M_3 ignoring the overall proportion of missing data, and only 2.3 % of the variance of the bias is explained if $Cor(\xi, \theta)$ and the overall proportion of missing data are excluded from the model. The results confirm that the degree of nonignorability given by $Cor(\xi, \theta)$ and the overall proportion of missing data are the most important factors that determine the bias of ML person parameter estimates. Note that generalizability is limited in the simulation study. In real applications many other factors that were not considered here may contribute to the bias.

Bias of Warm's WML person parameter estimates Lord (1983b) described the bias of ML estimates in tests consisting of a finite number of items. Warm (1989) proposed a

Table 3.4: Determination Coefficients R^2 of Saturated Regression Models $M_1 - M_4$ for Mean Biases of IRT Person and Item Parameter Estimates.

Dependent variable	$R^2_{M_1}$	$R^2_{M_2}$	$R^2_{M_3}$	$R^2_{M_4}$
Mean $Bias(\hat{\xi}_{ML})$	0.389^{***}	0.107^{***}	0.269^{***}	0.024^{***}
Mean $Bias(\hat{\xi}_{WML})$	0.410^{***}	0.193^{***}	0.240^{***}	0.056^{***}
Mean $Bias(\hat{\xi}_{EAP})$	0.019	/	/	/
Mean $Bias(\hat{\beta}_i)$	0.416^{***}	0.121^{***}	0.261^{***}	0.023^{***}
Mean $Bias(\hat{\alpha}_i)$	0.051^{***}	0.035^{***}	0.027^{***}	0.022^{***}

weighted ML (WML) estimator that reduces the bias of traditional ML estimates. Many authors found that the WML estimator should be preferred to traditional ML estimates (e. g. Hoijtink & Boomsma, 1996). Warm (1989) suggested to weight the likelihood function by the square root $\sqrt{I(\xi)}$ of the item information function $I(\xi)$. Hence, the weighted ML estimate $\hat{\xi}_{WML}$ is that value of Ω_ξ that maximizes the weighted pattern likelihood $\mathcal{L}^{(w)}(\boldsymbol{y_n}; \boldsymbol{\iota}) = \mathcal{L}(\boldsymbol{y_n}; \boldsymbol{\iota}) \cdot \sqrt{I(\xi)}$. Hence

$$\mathcal{L}^{(w)}(\boldsymbol{y_n}; \boldsymbol{\iota}) = P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \xi; \boldsymbol{\iota}) \sqrt{I(\xi)} \tag{3.51}$$

$$= \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota}) \sqrt{I(\boldsymbol{\xi})}. \tag{3.52}$$

The first derivative of the weighted log-likelihood function $\ell^{(w)}(\boldsymbol{y_n}; \boldsymbol{\iota})$ is

$$\frac{\partial}{\partial \xi} \ln[\ell^{(w)}(\boldsymbol{y_n}; \boldsymbol{\iota})] = \sum_{i=1}^{I} \alpha_i [y_i - P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})] + \frac{\partial}{\partial \xi} \ln\Big(\sqrt{I(\xi)} \Big). \tag{3.53}$$

In the case of the 2PLM for dichotomous items Y_i, the second summand in Equation 3.53 is

$$\frac{\partial}{\partial \xi} \ln\Big(\sqrt{I(\xi)} \Big) = \frac{1}{2I(\xi)} \sum_{i=1}^{I} \alpha_i^3 P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})^2 P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota}). \tag{3.54}$$

Setting Equation 3.53 to zero and solving for ξ yields the weighted ML estimate $\hat{\xi}_{WML}$. If any missing data mechanism exists, only observed item responses can be used for person parameter estimation . Hence, the weighted response pattern likelihood $\mathcal{L}^{(w)}(\boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ of the observed items is maximized and only the information of the observed items $I_{obs}(\xi)$ is involved. Hence, under any missing data mechanism the pattern likelihood of the ob-

served items is

$$\mathcal{L}^{(w)}(\boldsymbol{y_{n;obs}};\boldsymbol{\iota}) = P(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \xi;\boldsymbol{\iota}) \sqrt{I_{obs}(\xi)} \tag{3.55}$$

$$= \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \xi;\boldsymbol{\iota})^{D_i=d_i} \sqrt{I_{obs}(\xi)}. \tag{3.56}$$

Accordingly, the first derivative of the logarithm of $\mathcal{L}^{(w)}(\boldsymbol{y_{n;obs}};\boldsymbol{\iota})$ is

$$\frac{\partial}{\partial\xi} \ln[\ell^{(w)}(\boldsymbol{y_{n;obs}};\boldsymbol{\iota})] = \sum_{i=1}^{I} d_i\alpha_i[y_i - P(Y_{ni} = 1 \,|\, \xi;\boldsymbol{\iota})] + \frac{\partial}{\partial\xi} \ln\Big(\sqrt{I_{obs}(\xi)}\Big), \tag{3.57}$$

with

$$\frac{\partial}{\partial\xi} \ln\Big(\sqrt{I_{obs}(\xi)}\Big) = \frac{1}{2I_{obs}(\xi)} \sum_{i=1}^{I} d_i\alpha_i^3 P(Y_{ni} = 1 \,|\, \xi;\boldsymbol{\iota})^2 P(Y_{ni} = 0 \,|\, \xi;\boldsymbol{\iota}). \tag{3.58}$$

As in the case of the ML estimator, the estimation equation of the WML estimates consists also on the conditional probabilities $P(Y_{ni} = y_{ni} \,|\, \xi;\boldsymbol{\iota})$ and model parameters. Hence, the WML estimator is expected to be similarly affected by item nonresponses as the traditional ML estimate. This is the more since Warm proved that $\hat{\xi}_{ML}$ and $\hat{\xi}_{WML}$ are asymptotically equally distributed. Figure 3.9 shows the relationship between the bias of $\hat{\xi}_{WML}$ and the latent variable ξ as well as the number of non-responses (Data Example A). The bias is weakly correlated to the ability ($r = -0.061$, $t = -2.732$, $df = 1998$, $p = 0.006$). As Rost (2004) stated, it is a characteristic of the WML estimates that values at the lower end of ξ tend to be overestimated while those at the upper end tend to be underestimated. This shrinkage is typical for Bayesian estimators. Indeed, although WML is not considered to be a Bayesian estimator, Jeffrey (Jeffrey, 1961) proposed to use the square root of the information function as a non-informative prior distribution. Insofar, Warms' WML estimator can also be regarded as a Bayesian estimator (Held, 2008; Hoijtink & Boomsma, 1996). Hence, the negative correlation between the bias and the latent variable ξ may reflect the shrinkage effect rather than the bias due to item nonresponses since the latent ability and the number of missing items were substantially negatively correlated ($r = -0.719$). The bias of the WLM estimates and the number of non-responses were slightly positively correlated ($r = 0.063$, $t = 2.839$, $df = 1998$, $p = 0.005$). Also, the mean $\overline{\hat{\xi}}_{WML} = -0.088$ of the WML estimates is significantly different from zero ($t = -3.394$, $df = 1999$, $p < 0.001$) although the model was identified with $E(\xi) = 0$. The loss of information due to item nonresponses is reflected by a considerably reduced marginal reliability

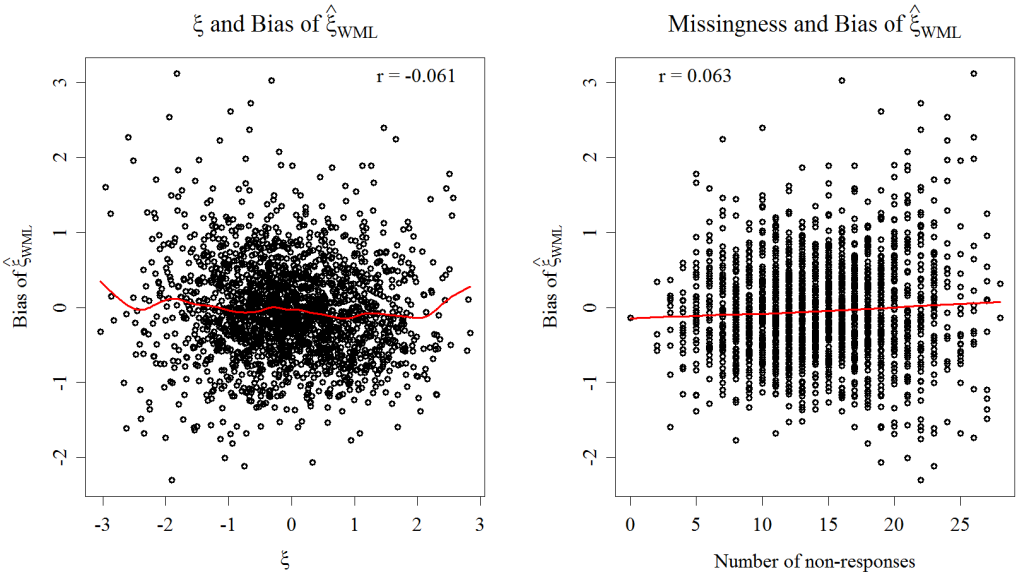

Figure 3.9: Relationship between the bias of the Warm's weighted ML estimates of Data Example A and the latent variable ξ (left) and the number of non-responses (right). The red line is a smoothing spline regression.

$Rel(\hat{\xi}_{WML}) = 0.614$ and a twofold higher mean squared error (MSE = 0.427) compared to the complete data (see Table 3.3).

On average the bias pattern of WML estimates (see Figure 3.10) found in the simulation study is rather similar to that of the ML estimate (cf. Figure 3.8). Again, the correlation $Cor(\xi, \theta)$ and the overall proportion of missing data seem to be the most influential factors of the WML bias. Both factors interact with one another. That is, no systematic bias could be found if the missing data mechanism was MCAR ($Cor(\xi, \theta) = 0$), even for large proportions of missing data. The higher the correlation $Cor(\xi, \theta)$ is, the more bias results from increasing proportions of missing data. 41.0 % of the variance in the bias could be explained by all factors in the models (see Table 3.4). This proportion dropped to 5.6 % if the correlation $Cor(\xi, \theta)$ and the overall proportion of missing data were not included in the regression. In Data Example A a small positive correlation between the bias and the number of non-responses was found. The simulation study confirmed that biasedness of WML estimates and the correlation between the latent ability and number of missing items depends on the correlation $Cor(\xi, \theta)$, the overall proportion of missing data, and correlation $r(\gamma, \beta)$ (see Figure 3.11). Complex interactions between these factors seem to exist. If $r(\gamma, \beta)$ is low, the bias of the WLM-estimates is increasingly positive with rising proportions of missing data and higher correlations $Cor(\xi, \theta)$. However, if $r(\beta, \gamma) = 0.5$

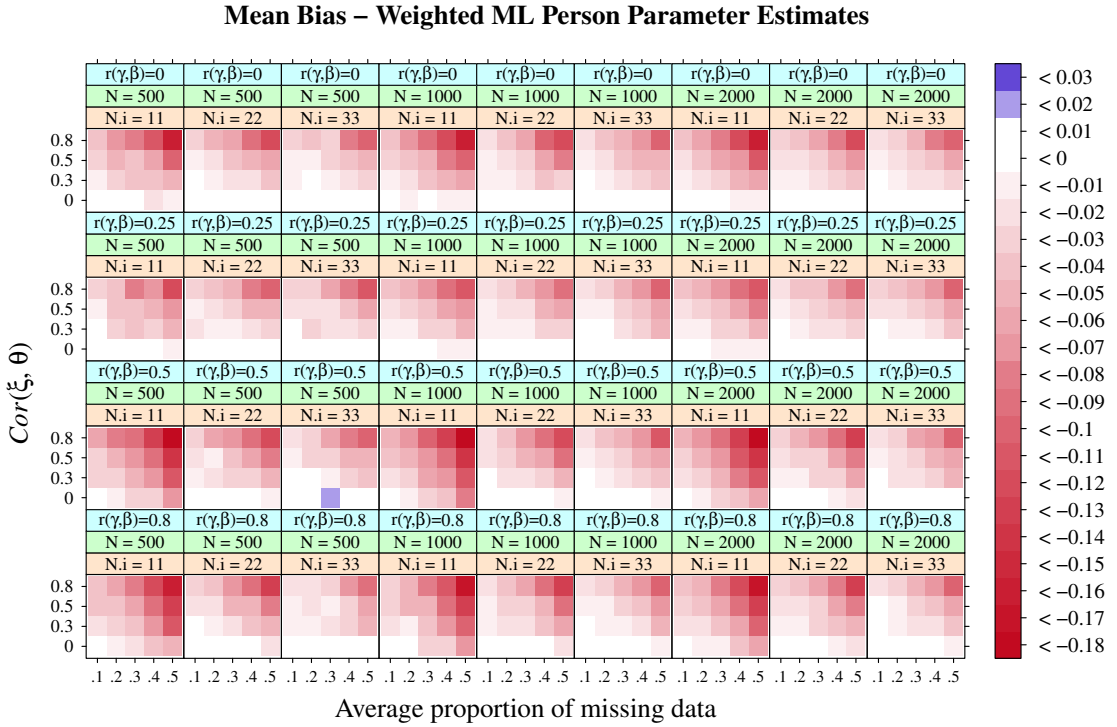

Figure 3.10: Mean bias of Warm's weighted ML person parameter estimates using the 1PLM (simulation study).

or $r(\beta, \gamma) = 0.8$, then a negative correlation between the bias of the WLM estimates and the number of nonresponses was found, particularly if the number of items is small and the missing data mechanism w.r.t. $\boldsymbol{Y}$ was MCAR ($Cor(\xi, \theta) = 0$). This is all the more interesting since the WLM estimator was on average unbiased if the nonresponse mechanism was MCAR (see Figure 3.10). The results suggest that the WML and traditional ML person parameter estimates are similarly affected by item nonresponses. Despite minor differences, both tend to be increasingly negatively biased with increasing proportions of missing data and higher correlations between persons' proficiency and their response propensity.

Bias of EAP person parameter estimates The expected a posteriori person parameter estimates $\hat{\xi}_{EAP}$, or simply EAPs, are Bayesian estimators. In the Bayesian framework the parameters are regarded as random variables with a distribution. Thus, the model parameters $\boldsymbol{\iota}$ and the manifest variables $\boldsymbol{Y}$ have a joint distribution $g(\boldsymbol{Y}, \boldsymbol{\iota})$ that can be factored into $g(\boldsymbol{Y}, \boldsymbol{\iota}) = g(\boldsymbol{Y} \mid \boldsymbol{\iota}) g(\boldsymbol{\iota})$ with $g(\boldsymbol{\iota})$ as the prior distribution. Using MML estimation the item

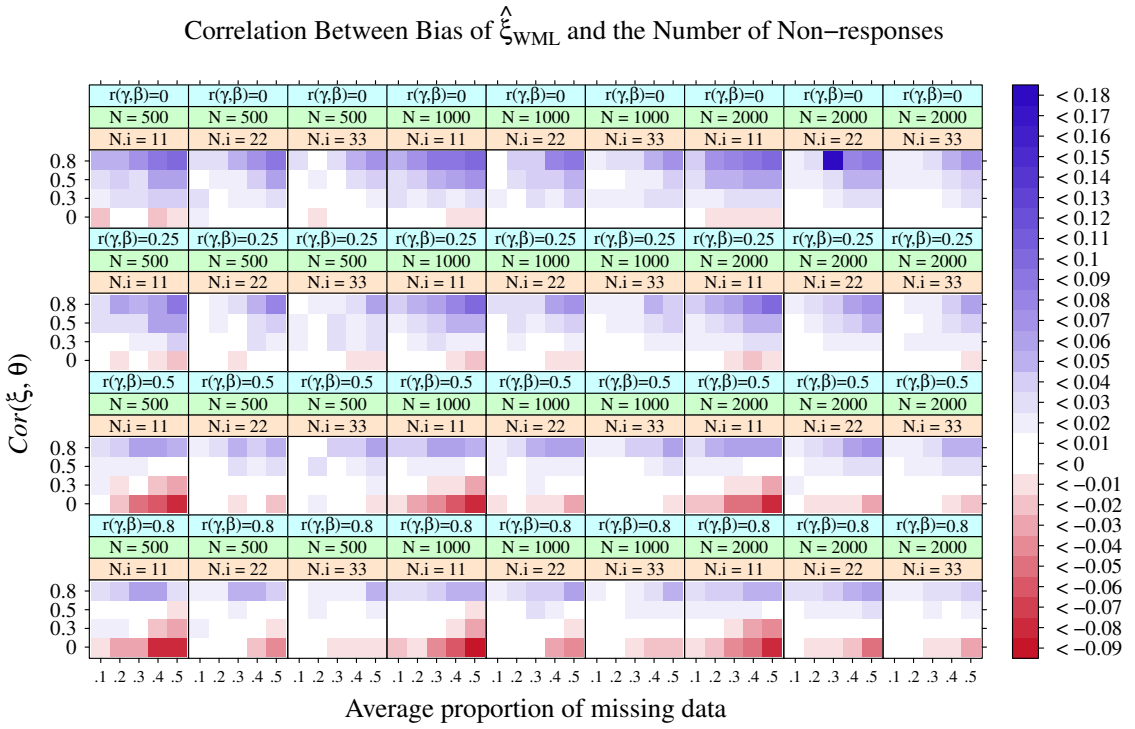

Figure 3.11: Correlation between the bias of Warm's weighted ML person parameter estimates and the number of non-responses (simulation study).

parameters are typically estimated first and then taken as fixed when estimating EAPs in a second step. Hence, the joint distribution of the item responses and the latent variables aimed to be estimated is $g(\boldsymbol{Y}, \xi; \boldsymbol{\iota}) = g(\boldsymbol{Y} \mid \xi; \boldsymbol{\iota}) g(\xi)$. In this case $\boldsymbol{\iota}$ consists of the item parameters and is replaced by the vector $\hat{\boldsymbol{\iota}}$ of sample estimates in real applications. The first factor is simply the conditional distribution $g(\boldsymbol{Y} \mid \xi; \boldsymbol{\iota}) = P(\boldsymbol{Y} = \boldsymbol{y} \mid \xi; \boldsymbol{\iota})$ that is also involved in ML and WML estimation. All Bayesian inferences rest upon the posterior distribution (e. g. Gelman et al., 2003; Held, 2008; Skrondal & Rabe-Hesketh, 2004). That is the distribution of the estimand given the observed data and researchers? prior belief expressed by the prior distribution. The posterior distribution of the latent variable ξ of a randomly chosen person n is

$$g(\xi \mid \boldsymbol{Y}_n = \boldsymbol{y}_n; \boldsymbol{\iota}) = \frac{P(\boldsymbol{Y}_n = \boldsymbol{y}_n \mid \xi; \boldsymbol{\iota}) g(\xi)}{\int_{\mathbb{R}} P(\boldsymbol{Y}_n = \boldsymbol{y}_n \mid \xi; \boldsymbol{\iota}) g(\xi) d\xi}, \tag{3.59}$$

given $\Omega_\xi = \mathbb{R}$. The EAP is defined as the expected value of the posterior distribution. In a unidimensional latent trait model this is

$$\hat{\xi}_{EAP} = \frac{\int_{\mathbb{R}} \xi \cdot P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \xi; \boldsymbol{\iota}) g(\xi) d\xi}{\int_{\mathbb{R}} P(\boldsymbol{Y_n} = \boldsymbol{y_n} \,|\, \xi; \boldsymbol{\iota}) g(\xi) d\xi}. \tag{3.60}$$

The denominator is simply the unconditional probability $P(\boldsymbol{Y_n} = \boldsymbol{y_n}; \boldsymbol{\iota})$ given a particular model indexed by $\boldsymbol{\iota}$. In the nominator the pattern likelihood (see Equation 3.47) is involved. Hence, Equation 3.60 can be written as

$$\hat{\xi}_{EAP} = \frac{\int_{\mathbb{R}} \xi \cdot \mathcal{L}(\boldsymbol{y_n}; \boldsymbol{\iota}) g(\xi) d\xi}{P(\boldsymbol{Y_n} = \boldsymbol{y_n}; \boldsymbol{\iota})}. \tag{3.61}$$

However, under any missing data mechanism the EAPs are estimated only due to the observed items $\boldsymbol{y_{obs}}$. The EAP estimator is then

$$\hat{\xi}_{EAP} = \frac{\int_{\mathbb{R}} \xi \cdot \mathcal{L}(\boldsymbol{y_{obs}}; \boldsymbol{\iota}) g(\xi) d\xi}{P(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota})}. \tag{3.62}$$

The formulas of the EAP person parameter estimates shows that again the item parameters are involved since the probabilities $P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota})$ are included. For this reason, the accuracy of EAPs depends also on the precision of item parameter estimates. In contrast to ML and WML estimates, the prior $g(\xi)$ is also influential. Generally, Bayesian estimates suffer from the so-called shrinkage effect. That is, the estimates tend toward the mean of the prior distribution. The shrinkage effect depends on the variance of the prior distribution and the amount of information given by observed data. The less information is available from observed data $\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}$, the more impact the prior distribution has in the calculation of $\hat{\xi}_{EAP}$ (e. g. Gelman et al., 2003; Held, 2008). If the number of answered items varies across test takers, then the shrinkage effect varies as well depending on amount of missing data. On average the shrinkage should be enhanced under any missing data mechanism resulting in a variance reduction of the EAP estimates. Indeed, as Table 3.3 shows, the variance of the EAPs with missing data is barely 0.632, compared to 0.859 of the complete data. It can also be seen that the MSE of the EAPs are the lowest compared to ML and WML estimates. However, the bias of the EAPs is correlated with ξ even in the complete data ($r = -0.378$). This is a side effect of the shrinkage effect which is considerably increased when missing data are present. In Data Example A the correlation between the bias of EAPs of the incomplete data and ξ increased to $r = -0.608$ (see also Figure 3.12). Furthermore, the missing data mechanism in Data Example A was

non-ignorable implied by $Cor(\xi, \theta) = 0.8$. The negative correlation between the bias and ξ on the one hand, and the positive correlation $Cor(\xi, \theta) = 0.8$ on the other hand, imply that the bias of EAPs should be positively correlated with the number of non-responses. Figure 3.12 confirms a substantial relationship between the bias and the number of non-responses. These results imply that, under certain conditions, the test takers may profit

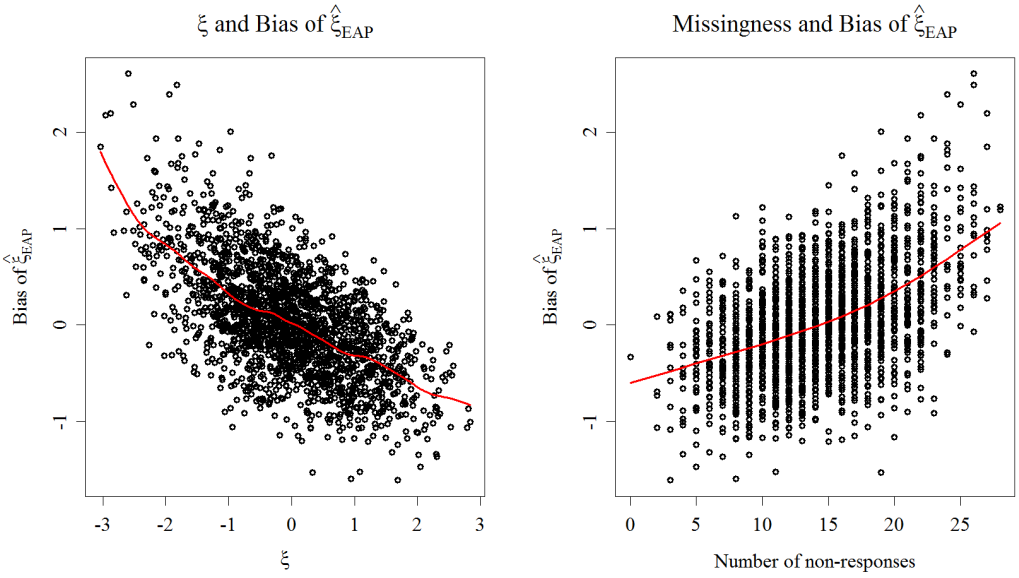

Figure 3.12: Relationship between the bias of the EAP person parameter estimates of Data Example A and the latent variable ξ (left) and the number of non-responses (right). The red line is a smoothing spline regression.

from omitting items. Especially persons with low ability levels profit from the shrinkage effect of the EAP estimator. In turn, highly proficient persons are affected adversely due to non-responses. In a single data set the number of non-responses varies across the test takers. Conclusively, the shrinkage effect varies as well. Here, it is argued that this undermines the comparability of the Bayesian point estimates such as the EAP. Compared with the ML and the WML estimators this seems to be a unique problem of Bayesian estimates.

In the simulation study most findings from Data Example A could be confirmed to be stable and systematic across considered conditions. As Figure 3.13 shows, the average bias of the EAPs is the lowest of all considered estimators in this work that is consistent with the lowest MSE and the lowest mean bias in Data Example A. Surprisingly, on average there is almost no systematic bias of the EAPs. This distinguishes EAPs from ML and WML estimates. However, there is a conditional bias given the latent variable

ξ due to the shrinkage effect and given the number of non-responses when $Cor(\xi, \theta) \neq 0$. As Figure 3.14 shows, the correlation between the bias of EAPs and the number of nonresponses is mainly driven by the $Cor(\xi, \theta)$. If the missing data mechanism is MCAR due to $Cor(\xi, \theta) = 0$, implying that ξ and the number of nonresponses are uncorrelated as well, the correlation of the EAP bias and the number of item nonresponses is close to zero. However, the higher the correlation $Cor(\xi, \theta)$, the stronger the negative correlation between the bias of the EAPs and the number of nonresponses is. Values of $r \approx 0.6$ are reached if $Cor(\xi, \theta) = 0.8$. Hence, from EAP scoring, especially, test takers with below-average proficiency levels would profit from skipping difficult items because the increased shrinkage effect results in higher scores closer to the mean of the prior distribution. In turn, persons with above-average abilities will not profit from omission of even difficult items, since the increasing shrinkage effect results in lower EAP estimates.

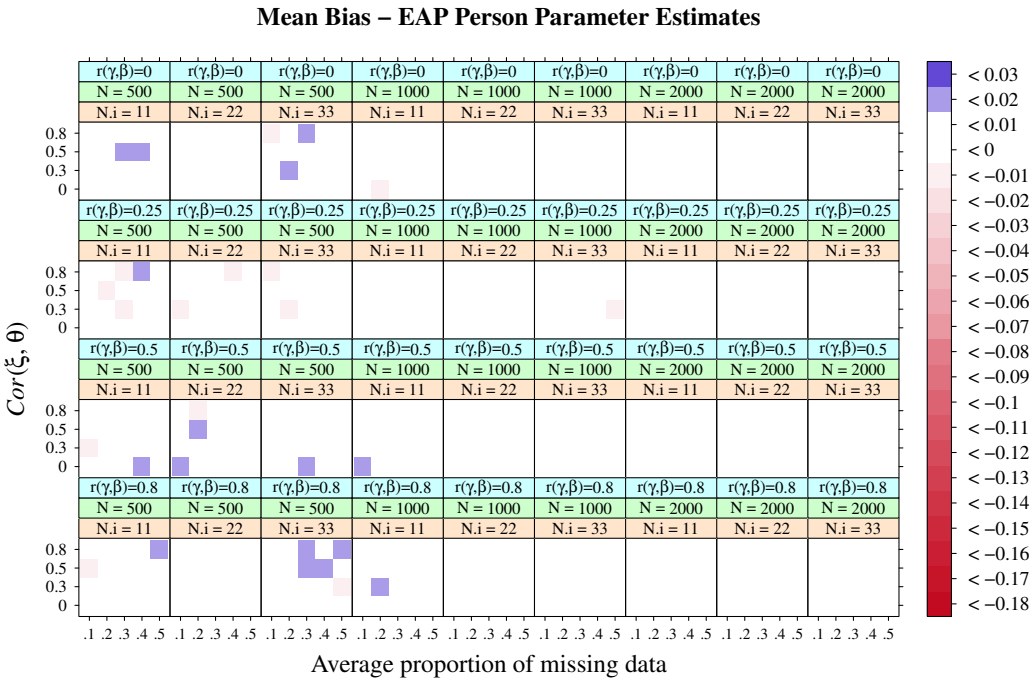

Figure 3.13: Mean bias of EAP person parameter estimates using the 2PLM (simulation study).

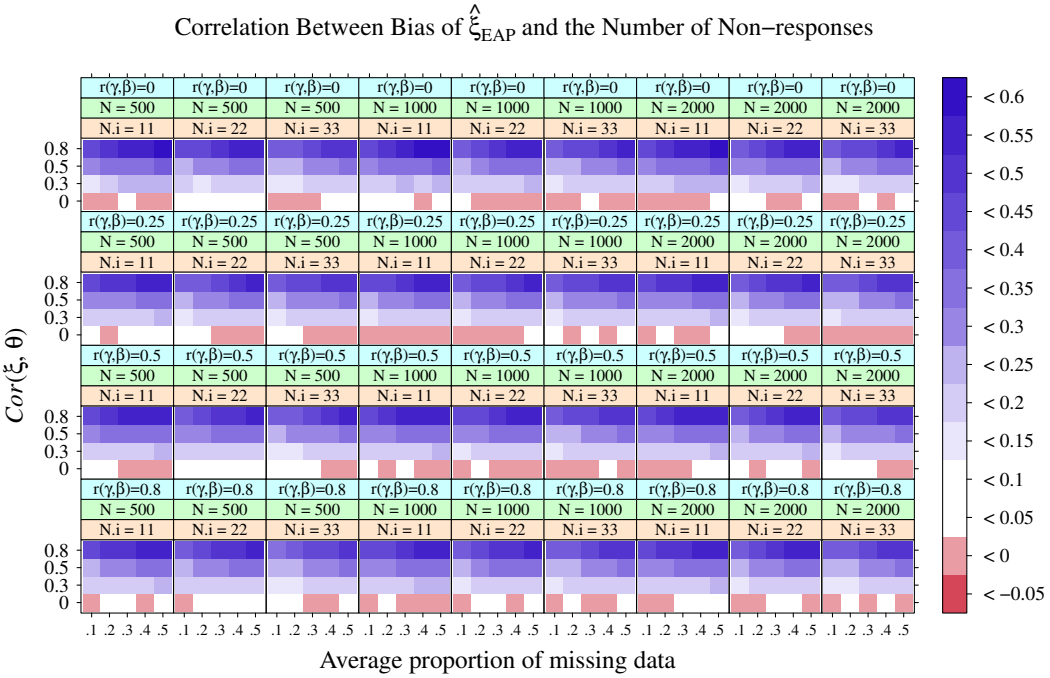

Figure 3.14: Mean correlation between bias of the EAP estimates and number of omitted responses (simulation study).

3.2 Item Parameter Estimates

3.2.1 Expected Values $E(Y_i)$

In CTT models the true score variables τ_i are linear functions of each other and, typically, linear functions of the latent variable ξ. Therefore, these models are mostly inappropriate for single categorical items Y_i [9]. For that reason , CTT models are commonly based on test scores, such as sum scores, of either complete tests or sub-tests (e. g. item parcels) instead of single items. Nevertheless, it is common in CTT to provide measures of difficulty with respect to single items Y_i that constitute the test. Typically, the unconditional expected values $E(Y_i)$ or conditional expected values $E(Y_i \,|\, \boldsymbol{Z} = \boldsymbol{z})$ are estimated by the sample item means $\bar{y}_i$ and $\bar{y}_{i|z}$ respectively. For categorical variables with K response categories

[9] There are some exceptions. For example the binomial model (Rost, 2004) for dichotomous items Y_i with equal item difficulties for all items allows for linearity.

the expected value of $E(Y_i)$ is the weighted sum

$$E(Y_i) = \sum_{y=0}^{K} y \cdot P(Y_i = y). \tag{3.63}$$

In the case of dichotomous items $E(Y_i)$ is simply

$$\begin{aligned} E(Y_i) &= 0 \cdot P(Y_i = 0) + 1 \cdot P(Y_i = 1) \\ &= P(Y_i = 1). \end{aligned} \tag{3.64}$$

Since the true scores τ_i are regressions of Y_i on the person variable U, equality $E(Y_i) = E(\tau_i)$ is implied (e. g. Steyer, 1989; Steyer & Eid, 2001; Steyer, 2002). In measurement models including a latent variable $\xi = f(U)$ and $\tau_i = f_i(\xi)$ implying that $\tau_i = (f_i \circ f)(U)$ the expected value is $E(Y_i) = E[E(Y_i \,|\, \xi)]$. If Y_i is dichotomous this is

$$\begin{aligned} E(Y_i) &= E[P(Y_i = 1 \,|\, \xi)] \\ &= \int_{\mathbb{R}} P(Y_i = 1 \,|\, \xi) g(\xi) d\xi \end{aligned} \tag{3.65}$$

Hence, the expected values of the items depend on the distribution of the latent variable ξ. That is why CTT based item difficulties are population specific measures. $E(Y_i)$ is not purely a measure of the items difficulty but a measure of the difficulty with respect to a particular population with a specific ability distribution of ξ. For that reason, several conditional difficulties $E(Y_i \,|\, \boldsymbol{Z} = z)$ in subpopulations given by $\boldsymbol{Z} = z$ can be estimated. Why is this important in the context of missing data problems in psychological and educational measurement? Consider the example where a representative sample has been drawn for an assessment. The test takers, however, are unwilling or unable to complete all items of the test. Hence, there are item nonresponses due to omitted or not-reached items. If the item means are computed only by the observed item responses, then the expected values $E(Y_i \,|\, D_i = 1)$ is estimated instead of $E(Y_i)$. The sample mean can also be regarded as a random variable $\bar{Y}_i$ with a sampling distribution. Under any missing data mechanism the item mean $\bar{Y}_{i;obs}$ of the observed responses can be written as

$$\bar{Y}_{i;obs} = \frac{\sum_{n=1}^{N} D_{ni} \cdot Y_{ni}}{\sum_{n=1}^{N} D_{ni}}. \tag{3.66}$$

If no missing data mechanism exists, then $D_{ni} = 1$ (for all $n = 1, \ldots, N$). In this case the nominator is simply $\sum_{n=1}^{N} D_{ni} \cdot Y_{ni} = \sum_{n=1}^{N} Y_{ni}$ and the denominator is $\sum_{n=1}^{N} D_{ni} =$

N, implying that $\bar{Y}_{i;obs} = \bar{Y}_i$. However, if $P(D_i = 1) < 1$, then the observable values y_i of Y_i are realizations from the conditional distribution $g(Y_i \mid D_i = 1)$ instead of $g(Y_i)$ and $\bar{Y}_{i;obs}$ will be a consistent estimator of $E(Y_i \mid D_i = 1)$ instead of $E(Y_i)$. In section 2.3 the implications of the different missing data mechanisms were scrutinized. If the missing data mechanism w.r.t. Y_i is MCAR then $g(Y_i \mid D_i = 1) = g(Y_i)$ (see Equation 2.38) implying equality $E(Y_i \mid D_i = 1) = E(Y_i)$ as well. In this case $\bar{Y}_{i;obs}$ is an unbiased estimator of $E(Y_i)$. Under any other missing data mechanism as defined here in this work $g(Y_i \mid D_i = 1) \neq g(Y_i)$. Hence,$E(Y_i \mid D_i = 1) \neq E(Y_i)$. The mean $\bar{Y}_{i;obs}$ will be an unbiased estimator of $E(Y_i \mid D_i = 1)$ instead of $E(Y_i)$. Furthermore, if measurement invariance of the manifest variables Y_i given D_i hold true, in the case of dichotomous items, inequality $g(\xi \mid D_i = 1) \neq g(\xi)$ of the distribution of the latent variable is implied (see Equation 2.61). The expected value $E(Y_i \mid D_i = 1)$ of an dichotomous item is given by

$$\begin{aligned} E(Y_i \mid D_i = 1) &= E[P(Y_i = 1 \mid \xi) \mid D_i = 1] \\ &= \int_{\mathbb{R}} P(Y_i = 1 \mid \xi) g(\xi \mid D_i = 1) d\xi. \end{aligned} \tag{3.67}$$

In other words if the missing data mechanism is not MCAR, then the observed values of Y_i are item responses given by test takers that are not representative with respect to the latent ability distribution. As previously noted, CTT-based item difficulties expressed by expected values of manifest items are only meaningful with respect to a particular population defined by its distribution of the latent variable ξ. Considering that the missing data mechanism can be different for each item Y_i, it is possible that the sample means $\bar{y}_{i;obs}$ are calculated based on different subsamples that are representative of different subpopulations in terms of the distributions of the latent variable. Formally this means that $g(\xi \mid D_i = 1) \neq g(\xi \mid D_j = 1)$.

This can be illustrated using Data Example A. The sample in this example is representative with respect to the ability distribution. That is, all simulated cases are generated by drawing values of the latent variable from unit normal distribution. However, there is a high proportion of missing data. Due to the correlation $Cor(\xi, \theta) = 0.8$, the missing data mechanism w.r.t. each variable Y_i is non-ignorable implying that $g(\xi \mid D_i = 1) \neq g(\xi)$. Additionally, Data Example A was generated in such a way that more difficult items are generally more likely omitted. Figure 3.15 shows the result of the item means (left) and the estimated distributions $g(\xi \mid D_i = 1)$ of each item Y_i (right). The left panel of Figure 3.15 compares the true means $1/N \sum_{n=1}^{N} P(Y_i = 1 \mid \xi)$ and the observed means $\bar{y}_{i;obs}$ computed based on Equation 3.66. Apparently, there is a systematic bias. The item means

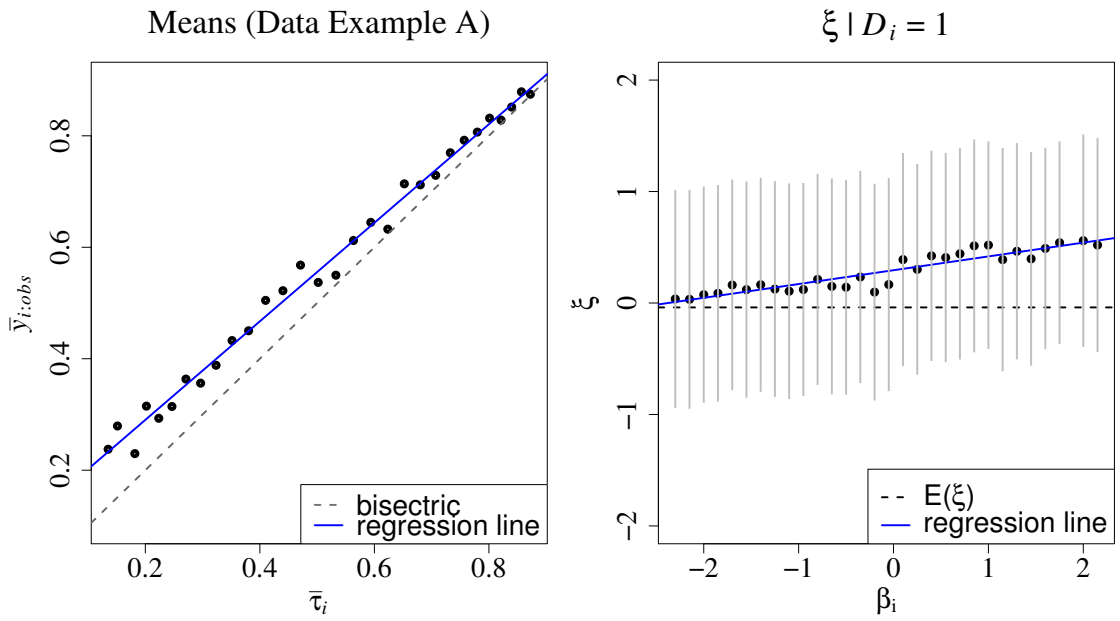

Figure 3.15: Means of the true scores and item means $\bar{y}_{i;obs}$ (right), and means and variances of $\xi \mid D_i = 1$ for each item (Data Example A).

are increasingly positively biased, the more difficult the item was. In conjunction with the theoretical considerations above, the right panel of Figure 3.15 makes clear why the bias increases depending on the item difficulty. The more difficult the items were, the higher the proportions of missing data were and the higher the average proficiency level of the responding test taker was. Recall that in Data Example A test takers differed in their mean test difficulties T_β depending on ξ. This item selection process is also reflected at the item level by the differences of the conditional distribution $g(\xi \mid D_i = 1)$ compared to the unconditional distribution $g(\xi)$. This example illustrates that a representative sample can become unrepresentative due to systematic missing data. The item means are estimates of item difficulties with respect to subpopulations that are potentially different across the items within a single test. Only if the missing data mechanism w.r.t. Y_i is MCAR, then $\bar{y}_{i;obs}$ will be an unbiased estimate of $E(Y_i)$. However, if the missing data mechanism w.r.t. Y_i is MAR given $\boldsymbol{Z}$, then equality $E(Y_i \mid \boldsymbol{Z}, D_i) = E(Y_i \mid \boldsymbol{Z})$ is implied from Equation 2.50. If $\boldsymbol{Z}$ is discrete, then the means of the observed item responses given the values $\boldsymbol{Z} = \boldsymbol{z}$ are unbiased estimators of $E(Y_i \mid \boldsymbol{Z} = \boldsymbol{z})$. This allows to compute adjusted means based on the regression $E(Y_i \mid \boldsymbol{Z})$ since $E[E(Y_i \mid \boldsymbol{Z}, D_i)] = E[E(Y_i \mid \boldsymbol{Z})] = E(Y_i)$. Hence, covariates can be used as auxiliary variables to yield unbiased item means if the missing data mechanism is MAR given $\boldsymbol{Z}$.

3.2.2 Threshold Parameters

In one-, two-, and three-parameter IRT models, threshold parameters describes the difficulty of an item. For dichotomous items the threshold β_i, or simply the item difficulty, is that value of ξ at which the probability $P(Y_i = 1 \,|\, \xi = \beta_i) = 0.5 + (c_i/2)$. c_i is the pseudo-guessing parameter of the three-parameter model. The Rasch and the Birnbaum models can be regarded as special cases of the 3PLM with $c_i = 0$ implying $P(Y_i = 1 \,|\, \xi = \beta_i) = P(Y_i = 0 \,|\, \xi = \beta_i) = 0.5$. The higher difficulty parameters β_i are, the more difficult the items Y_i are. The item difficulties and the latent variable ξ have a common metric. That is, β_i are locations on ξ. This is also true in multidimensional IRT models with a simple structure (between-item-dimensional MIRT models) and a subtractive parameterization, where the logit is $\alpha_i(\xi_m - \beta_i)$ for all items $i = 1, \ldots, I$. In within item-dimensional MIRT models the logit is $\sum_{m=1}^{M} \alpha_{im}\xi_m - \beta_i$. In this case the threshold parameters are not locations on a single latent dimension [10]. For simplicity here only the bias of item difficulty estimates $\hat{\beta}$ in unidimensional 1- and 2PL models is considered. The major advantage of parameters β_i as measures of item difficulties compared to expected values $E(Y_i)$ is their independence of the distribution of the latent variable ξ. Hence, IRT item parameters describe items' characteristics independently of a particular population. From this property it follows that item parameters can even be estimated unbiasedly if the sample of test takers is not representative with respect to the underlying ability distribution. Nevertheless, as demonstrated each item can be answered by a different subsample of respondents due to item nonresponses. In this case the item parameter estimates are potentially biased. Furthermore, since item difficulties are locations on the latent variable and ML and WML person parameter estimates were found systematically biased by non-ignorable missing data, estimates $\hat{\beta}_i$ may be biased as well. That applies all the more since the estimation equation involves also the person parameter estimates. The first derivative of the log-likelihood $\ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})$ of the observed data with respect to the item difficulties is

$$\frac{\partial \ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})}{\partial \beta_i} = -\alpha_i \sum_{n=1}^{N} d_{ni} \left[y_{ni} - P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota}) \right]. \tag{3.68}$$

If no missing data mechanism exists w.r.t. Y_i, $D_{ni} = 1$ for all $n = 1, \ldots, N$. Hence, the response indicators D_i can be omitted. In this case Equation 3.68 is a derivation of the

[10] Thus, even in within-item-dimensional MIRT models a multidimensional item difficulty can be constructed. Reckase (1985) proposed the distance between the origin of the multidimensional latent person parameter space to the point of maximum slope in the multidimensional item response surface (de Ayala, 2009; Reckase, 2009).

complete data likelihood since $\boldsymbol{Y} = \boldsymbol{Y_{obs}}$. In order to estimate β_i Equation 3.68 is set equal to zero. Since no closed-form expression exists, ML estimators are found iteratively by means of numerical methods. Using MML estimation the estimation equation of β_i is slightly different. The integral over the distribution of the latent variable ξ is involved. That is

$$\frac{\partial \ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})}{\partial \beta_i} = -\alpha_i \sum_{n=1}^{N} d_{ni} \int_{\mathbb{R}} \left[y_{ni} - P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota}) \right] g(\xi \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota}) d\xi. \tag{3.69}$$

. To reduce computational burdens due to numerical integration over the latent variable, the distribution $g(\xi)$ is replaced by a quadrature distribution $g(\xi_q)$ with Q values ξ_q (e. g. Baker & Kim, 2004). Hence, the continuous latent variables are discretized and the integral in Equation 3.70 becomes a sum over the conditional quadrature distributions $g(\xi_q \,|\, \boldsymbol{Y_n} = \boldsymbol{y_n}; \hat{\boldsymbol{\iota}})$. Although MML does not require the estimation of individual values of the latent variable, the conditional probabilities $P(\xi_q \,|\, \boldsymbol{Y_n} = \boldsymbol{y_n}; \hat{\boldsymbol{\iota}})$ that test taker n has the trait level ξ_q need to be estimated in the E-step. This calculation is required for each test taker with respect to each quadrature point. Finally, the estimation equation can be written as

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})}{\partial \beta_i} \quad = \quad -\alpha_i \Bigg[& \sum_{n=1}^{N} d_{ni} \sum_{q=1}^{Q} y_{ni} P(\xi_q \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota}) - \\ & \sum_{n=1}^{N} d_{ni} \sum_{q=1}^{Q} P(Y_{ni} = 1 \,|\, \xi_q; \boldsymbol{\iota}) P(\xi_q \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota}) \Bigg]. \end{aligned} \tag{3.70}$$

The minuend is the expected number of correct answers assuming a specified latent distribution $g(\xi)$ approximated by $g(\xi_q)$. The subtrahend is the expected number of correct answers given the same distributional assumption *and* the specified IRT model. Equation 3.70 illustrates why the prediction of the bias of IRT item parameters due to item nonresponses is so difficult. Both terms - the minuend and the subtrahend - involve quantities that depend on unknown model parameters indexed by $\boldsymbol{\iota}$. Even the calculation of the conditional probabilities $P(\xi_q \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ is affected by item parameters (e. g. Baker & Kim, 2004). Using the EM algorithm the expected numbers of correct answers in Equation 3.70 are calculated in the E-step using starting values or provisional estimates $\hat{\iota}$. In the M-step the updated estimates $\hat{\boldsymbol{\iota}}$ are computed, which are used again in the subsequent E-step. This cycle is repeated until a previously specified convergence criterion is reached. However, due to item nonresponses $\hat{\iota}$ can be biased resulting in biased es-

timates of conditional probabilities $P(Y_{ni} = 1 \mid \xi_q; \boldsymbol{\iota})$ as well as $P(\xi_q \mid \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$. These biases, in turn, result in potentially biased estimates of $\hat{\boldsymbol{\iota}}$ in the subsequent and final iteration step after convergence. Furthermore, the estimation of $P(\xi_q \mid \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ in the E-step depends not only on the observed item responses to item i but on all observed item responses provided by test taker n. If preferably easy items are answered with higher probabilities to be solved while difficult items are skipped, then these probabilities are potentially estimated with a systematic bias even if provisional estimates $\hat{\boldsymbol{\iota}}$ are unbiased.

Hence, although a clear prediction about biasedness of $\hat{\beta}_i$ is difficult, it is most likely that especially estimates $\hat{\beta}_i$ of difficult items in Data Example A will be negatively biased. The reason is that increasingly difficult items are answered by on average more proficient persons. Hence, the items seem to be easier than they really are. In other words, corresponding to the positive bias found in item means $\bar{y}_{i;obs}$, the estimates $\hat{\beta}_i$ are expected to be negatively biased. Note that the expected values $E(Y_i)$ are actually measures of item easiness instead of item difficulty. Therefore, the IRT item difficulty estimates are expected to be underestimated instead of overestimated.

Figure 3.16 compares the item difficulty estimates of Data Example A with the true item difficulties used for data simulation. For reasons of comparison, the difficulty estimates of the complete data are shown as well in the left graph, and the estimates of the incomplete data are depicted in the right graph. The estimates of the complete data are practically unbiased. The estimates resulting from the incomplete data reveal the expected pattern of the bias. The slope of the linear regression of the estimates on the true item difficulties is 0.934. That is significantly different from one (SE= 0.017, $t = -3.700$, $p < 0.001$) indicating a systematic bias. Especially the more difficult items are increasingly underestimated. However, the bias is small compared to the item means $\bar{y}_{i;obs}$. In fact, Rose et al. (2010) found as well that item parameter estimates are pretty robust even if the missing data mechanism is MNAR. However, the results of the simulation study revealed that the bias is systematically related to the missing data mechanism. As Figure 3.17 shows the pattern of biases is very close to that of the ML and WML person parameter estimates (cf. Figures 3.10 and 3.8). Generally the item difficulties tended to be underestimated. Exceptionally in the case of small sample sizes of $N = 500$, positive biases occurred as well with a nonsystematic pattern. The negative bias increases with stronger correlations $Cor(\xi, \theta)$. This effect is moderated by increasing overall proportions of missing data. The similarity of the biases of item difficulties and ML and WML person parameter estimates suggests that they are related. Using biased item difficulty estimates will most likely result in biased ML and WML person parameter estimation.

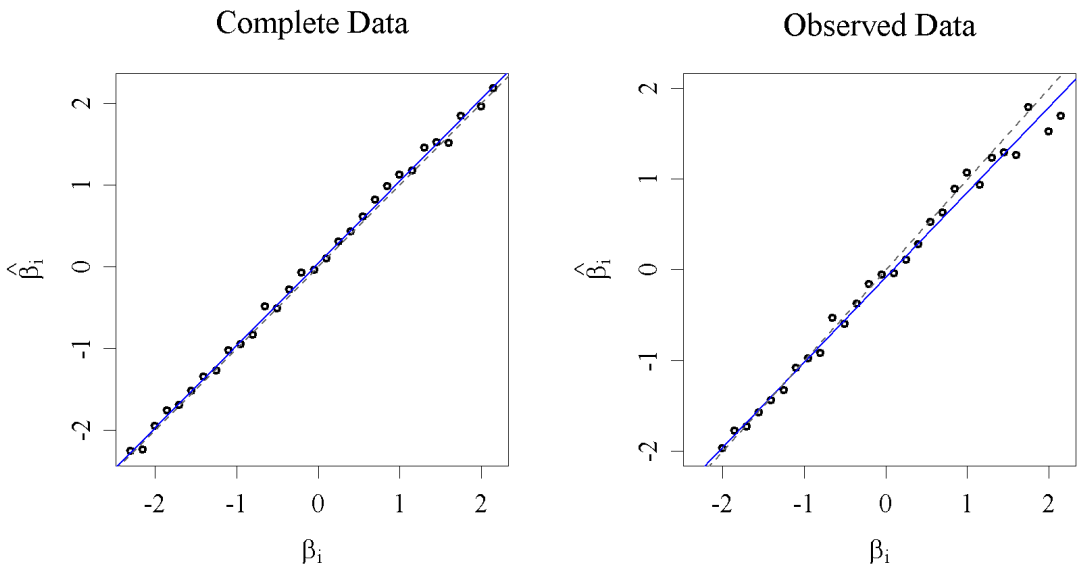

Figure 3.16: Comparison of true and estimated item difficulties using complete (left) and incomplete data (right) (Data Example A). The grey line is the bisectric. The blue line represents the regression line.

3.2.3 Item Discriminations

Finally, the impact of missing data on sample based estimates of item discrimination parameters $\alpha_i$in the Birnbaum model (Birnbaum, 1968) is studied. Again, neither closed form expressions nor sufficient statistics exist for estimation of α_i. ML estimation requires iterative methods. The first derivative of the log-likelihood $\ell(\boldsymbol{y}; \boldsymbol{\iota})$ with respect to α_i is involved. For item i that is

$$\frac{\partial \ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})}{\partial \alpha_i} = \sum_{n=1}^{N} d_{ni}(\xi - \beta_i) \left[y_{ni} - P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota}) \right] \tag{3.71}$$

Using MML estimation the first derivatives of the log-likelihood with respect to α_i is

$$\frac{\partial \ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})}{\partial \alpha_i} = \sum_{n=1}^{N} d_{ni} \int_{\mathbb{R}} (\xi - \beta_i) \left[y_{ni} - P(Y_{ni} = y_{ni} \,|\, \xi; \boldsymbol{\iota}) \right] g(\xi \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota}) d\xi. \tag{3.72}$$

As discussed, for estimation of item difficulties $g(\xi)$ is typically approximated by a quadrature distribution $g(\xi_q)$ to make numerical integration feasible. The estimation equation

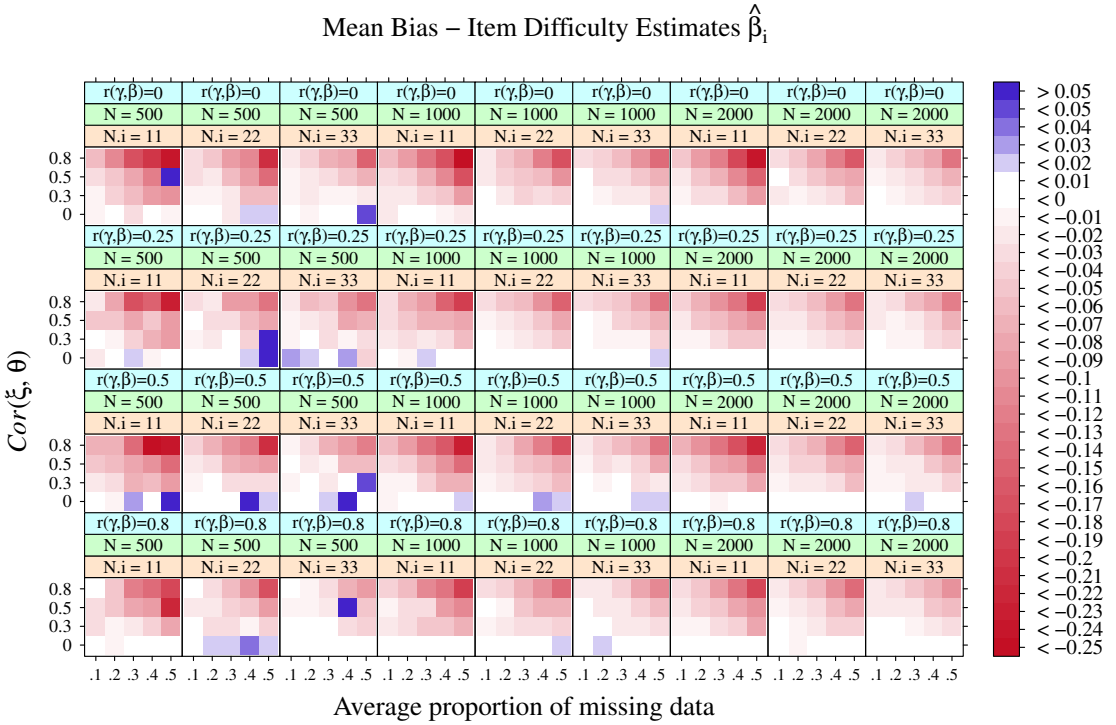

Figure 3.17: Mean bias of estimated item difficulties (simulation study).

using any discrete quadrature distribution $g(\xi_q)$ can be written as

$$\frac{\partial \ell(\boldsymbol{y_{obs}}; \boldsymbol{\iota})}{\partial \alpha_i} = \Big[\sum_{n=1}^{N} d_{ni} \sum_{q=1}^{Q} (\xi_q - \beta_i) y_{ni} P(\xi_q \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota}) - \sum_{n=1}^{N} d_{ni} \sum_{q=1}^{Q} (\xi_q - \beta_i) P(Y_{ni} = 1 \,|\, \xi_q; \boldsymbol{\iota}) P(\xi_q \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota})\Big]. \tag{3.73}$$

This estimation equation is similar to that of the item difficulties. Again, the conditional probabilities $P(\xi_q \,|\, \boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ to have a latent trait level ξ_q given the observed responses $\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}$ are involved, and the conditional probabilities $P(Y_{ni} = 1 \,|\, \xi_q; \boldsymbol{\iota})$ to solve item i given the latent ability is equal to trait level ξ_q. Equation 3.73 highlights that the prediction of the bias of the discrimination estimates is difficult to predict. For this reason biasedness was studied empirically. In Data Example A, the estimated discrimination parameters were found to be dependent on item difficulties even if the complete data were used for parameter estimation (left graph of Figure 3.18). This was also found for estimates $\hat{\alpha}_i$ obtained from incomplete data. The mean bias across the 30 items was not

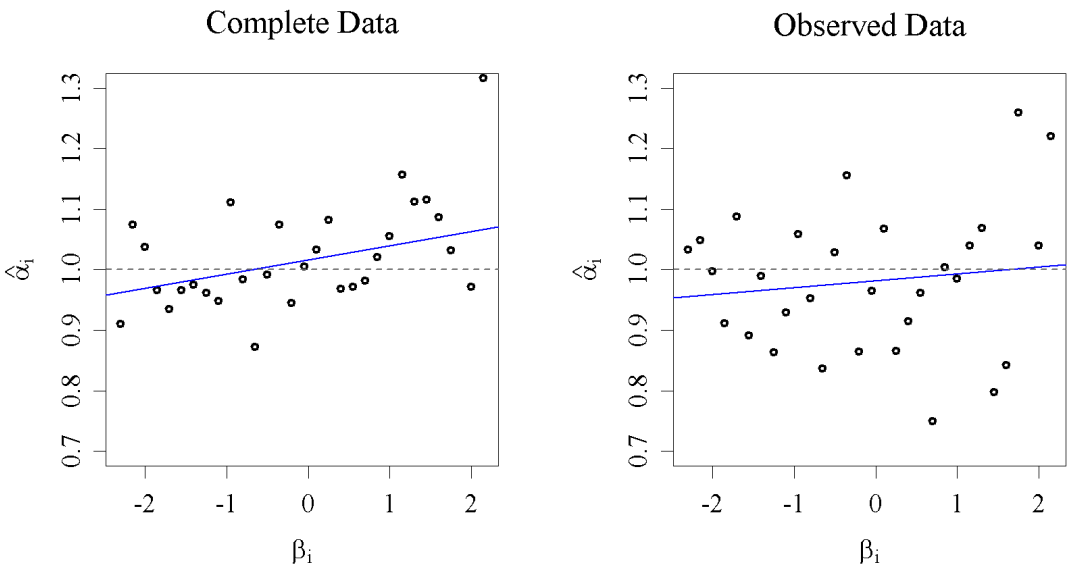

Figure 3.18: Estimated item discriminations using complete (left) and incomplete data (right) given the true item difficulties (Data Example A). The grey line is the bisectric. The blue line denotes the regression line.

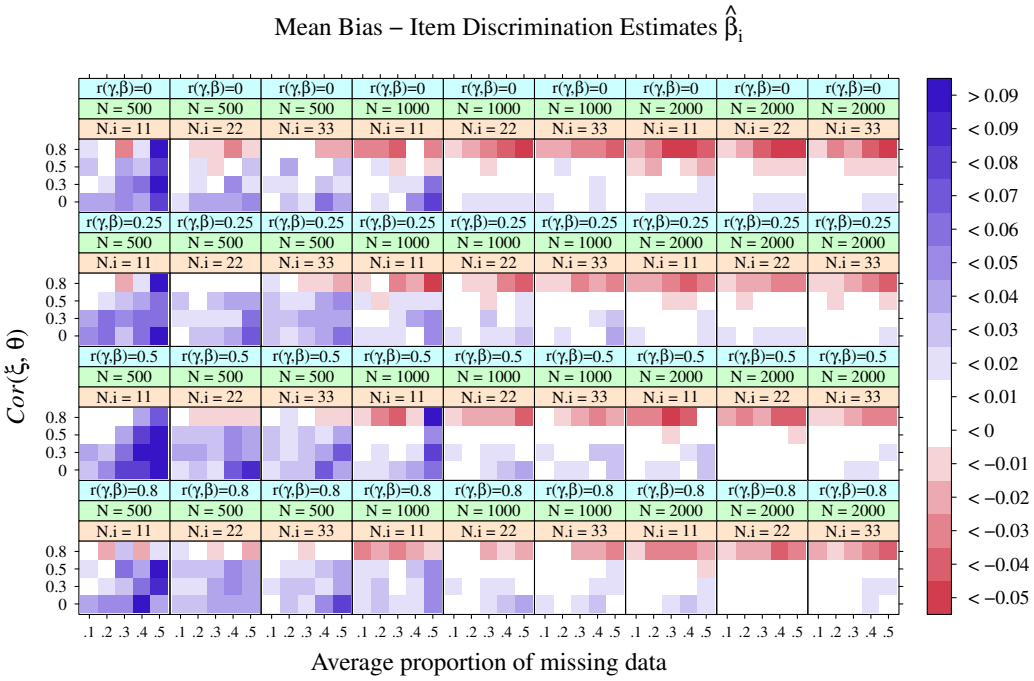

Figure 3.19: Mean bias of estimated item discriminations in the 2PLM (simulation study).

significantly different from zero using both the complete and the incomplete data. However, the variability of discrimination estimates is higher when incomplete data were used for parameter estimation (MSE = 0.014) compared to complete data (MSE = 0.008). In the simulation study there was also no evidence for a systematic bias due to item nonresponses (Figure 3.19). For a small sample size of $N = 500$ the item discrimination tends to be overestimated especially when the number of variables is low and the proportion of missing data is high. In sample sizes $N = 1000$ and $N = 2000$ a consistent positive bias of $\hat{\alpha}_i$ was found if the correlation between the latent ability and the latent response propensity was high $Cor(\xi, \theta) = 0.8$. However, as Table 3.4 shows, all chosen factors varied in the simulation study explained about 5 % of the variance in the mean bias of item discrimination estimates. Furthermore, in contrast to the bias of estimates $\hat{\beta}_i$, the correlation $Cor(\xi, \theta)$ and overall proportion of missing data was of minor importance. The sample size and the number of items i in the measurement model seem to have more impact. Indeed, a saturated regression model leaving out these two factors explains only 0.5 %. Hence, item discrimination parameters seem much less systematically biased due to item nonresponses than estimates of item difficulties and ML and WML person parameter estimates.

3.3 Standard Error Function and Marginal Reliability

Missing data are associated with a loss of information and are therefore expected to result in larger standard errors. In IRT models the standard errors of person parameter estimates are functions of the latent variables. The functional form of the standard error function $SE(\xi)$ of a unidimensional latent variable ξ is determined by the item parameters $\boldsymbol{\iota}$. Generally, the standard error function is $SE(\xi) = \sqrt{\boldsymbol{I}(\xi)}^{-1}$, with $\boldsymbol{I}(\xi)$ the test information function that is given by the sum of the item information functions $I_i(\xi)$ (e. g. de Ayala, 2009; Embretson & Reise, 2000). Hence, the standard error function can be written as

$$SE(\xi) = \left(\sqrt{\sum_{i=1}^{I} I_i(\xi)} \right)^{-1} \tag{3.74}$$

The item information functions $I_i(\xi) = \alpha_i^2 Var(Y_i \mid \xi)$, with the conditional variance $Var(Y_i \mid \xi) = P(Y_i = 1 \mid \xi; \alpha_i, \beta_i) P(Y_i = 0 \mid \xi; \alpha_i, \beta_i)$. Thus, the accuracy of the estimation of ξ by a given test depends solely on the item parameters α_i and β_i. However, if a nonresponse mechanism exists, then test takers select items randomly or systematically resulting in lost information and, thus, in larger standard errors. The standard error $SE_{obs}(\xi)$ function

given any missing data mechanism as defined above can be expressed using the response indicator variables D_i:

$$\begin{aligned} SE_{obs}(\xi) &= \left(\sqrt{\boldsymbol{I}_{obs}(\xi)}\right)^{-1} \\ &= \left(\sqrt{\sum_{i=1}^{I} D_i I_i(\xi)}\right)^{-1} \end{aligned} \tag{3.75}$$

Note that $SE_{obs}(\xi)$ is only defined if at least one item is observable. $SE_{obs}(\xi)$ is based on the observed item responses $\boldsymbol{y_{obs}}$. The missing pattern $\boldsymbol{D}$ is a random variables. Hence, there is not a single standard error function but as many standard error functions as response patterns exist minus one[11]. That is $I^2 - 1$. Figure 3.20 shows the estimated standard errors of different person parameter estimates in Data Example A. The blue line is the estimated standard error function of the complete data without missing values. The black dots are the standard errors for each simulated case with missing data, and the red line gives the average standard error for each value ξ across the observed missing pattern approximated by cubic smoothing spline. Figure 3.20 suggests that in presence of missing data the standard error function is not simply a function $f(\xi)$ of the latent variable but rather a function $f(\xi, \boldsymbol{D})$ of the latent variable and the missing pattern. Each missing pattern is associated with a different item subset that is completed by an individual test taker. As previously noted, test takers create their own test due to omissions of items or not completing the whole test in time. Each item subset can be regarded as a subtest with its own test information function and standard error function. The mean standard error function (red lines in Figure 3.20) is an estimator of the expected standard error of the latent variable ξ for a randomly drawn missing pattern. As expected, the standard errors are larger in presence of missing data compared to standard error that result from complete data. This increases the marginal error variance $Var(\varepsilon_{\hat{\xi}})$ as well and, therefore, the marginal reliability $Rel(\hat{\xi})$. Generally, the marginal reliability quantifies the accuracy of person parameter estimation by a single standardized coefficient, so that $0 \leq Rel(\hat{\xi}) \leq 1$. However, the standard error function $SE_(\xi)$ expresses that the accuracy of person parameter estimation depends on the latent variable. Therefore, the marginal reliability depends on the distribution of the latent variable and can be regarded as an average accuracy across the latent variable (de Ayala, 2009). Different marginal reliability coefficients have been proposed for different estimators (e. g. Andrich, 1988; Bock & Mislevy, 1982; Wright & Stone, 1979). Here the Andrich reliability is considered for ML and WML estimates, and the marginal EAP

[11] If $\boldsymbol{D} = \boldsymbol{0}$, then the standard error function is not defined.

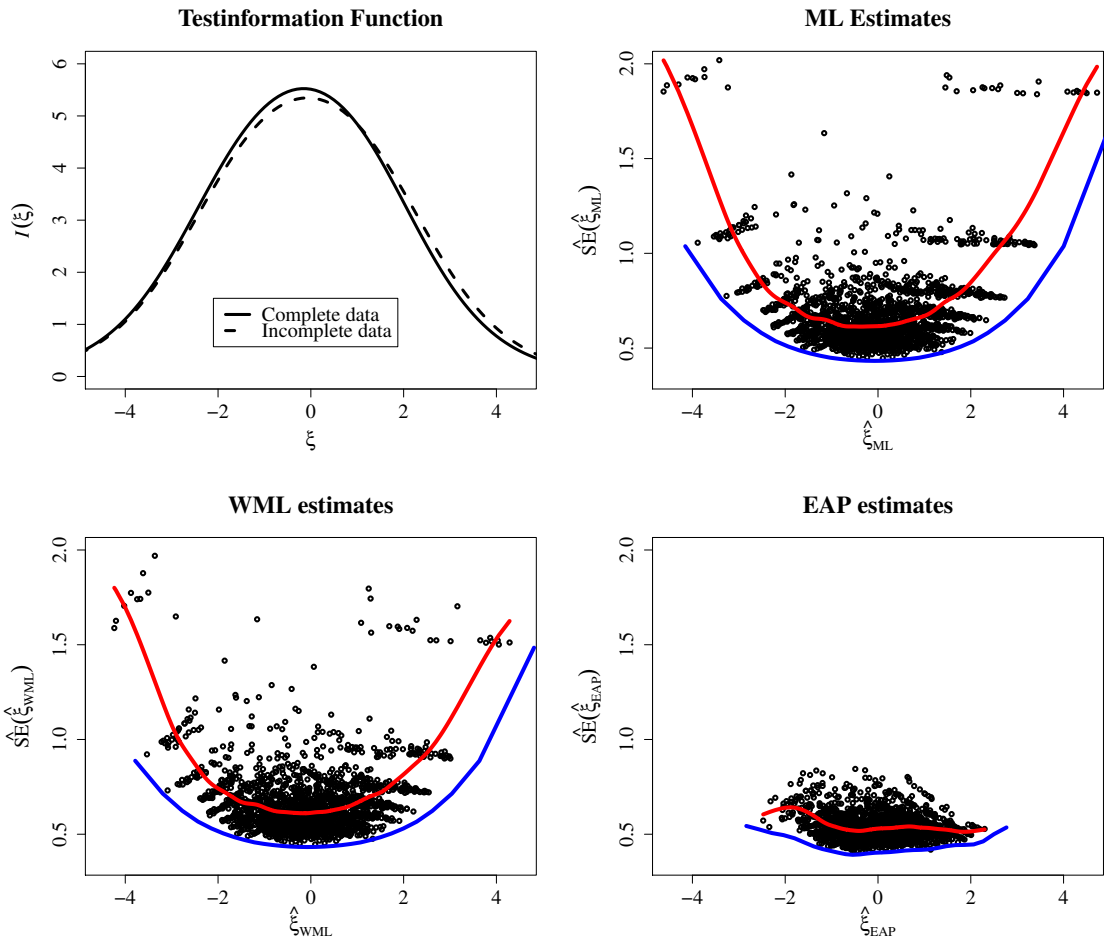

Figure 3.20: Model-implied test information functions (upper-left) and standard error functions (blue lines) based on item parameter estimates. The black dots represent ML-, WML- and EAP point estimates and their standard errors obtained from incomplete data (Data Example A). The red line approximates the mean standard errors.

reliability for EAP estimates. Andrich's reliability is defined as

$$Rel^{(A)}(\hat{\xi}) \quad = \quad 1 - \frac{Var(\varepsilon_{\hat{\xi}})}{Var(\hat{\xi})}, \tag{3.76}$$

with $\varepsilon_{\hat{\xi}} = \hat{\xi} - \xi$ the measurement error. Since the variance of the measurement error varies depending on the estimand ξ, the marginal error variance is the expected value of the error variance: $Var(\varepsilon_{\hat{\xi}}) = E(E[(\hat{\xi} - \xi)^2 \,|\, \xi]) = E[Var(\varepsilon_{\hat{\xi}} \,|\, \xi)]$. The conditional error variance $Var(\varepsilon_{\hat{\xi}} \,|\, \xi)$ is the squared standard error function $SE(\hat{\xi})^2$. In real applications, the marginal error variance is estimated by the mean of the squared standard errors over all test takers

in the sample. Hence, sample based estimate of Andrich's reliability can be written as

$$\widehat{Rel}^{(A)}(\hat{\xi}) \quad = \quad 1 - \frac{\frac{1}{N}\sum_{n=1}^{N}\widehat{SE}(\hat{\xi}_n)^2}{s^2(\hat{\xi})}. \tag{3.77}$$

This equation reveals that the sample-based estimate of the Andrich reliability is potentially affected by missing data in different ways. At first, person and item parameter estimates are involved. It was shown previously that biased item parameter estimates can result in biased person parameter estimates. Furthermore, the test information and standard error functions are potentially biased due to biased item parameter estimates. As the upper left graph of Figure 3.20 shows, only small differences between the test information functions estimated by item parameter estimates of complete and incomplete data were found in Data Example A. However, in the beginning of this section it was shown that in presence of missing data the standard error function is no longer a function of ξ alone, but a function $f(\xi, \boldsymbol{D})$ of the latent variable and the response indicator vector. Accordingly, the Andrich's reliability in presence of any missing data mechanism as defined in Section 2.2 is

$$\widehat{Rel}^{(A)}_{obs}(\hat{\xi}) \quad = \quad 1 - \frac{\frac{1}{N}\sum_{n=1}^{N}\widehat{SE}_{obs}(\hat{\xi}_n)^2}{s^2(\hat{\xi})}. \tag{3.78}$$

The Equations 3.77 and 3.78 seems to be almost identical. However, conceptually there is an important difference. The estimated marginal error variance without a missing data mechanism is $Var(\varepsilon_{\hat{\xi}}) = E[Var(\varepsilon_{\hat{\xi}} \,|\, \xi)]$, which is different from $Var(\varepsilon_{\hat{\xi}}) = E[Var(\varepsilon_{\hat{\xi}} \,|\, \xi, \boldsymbol{D})]$ if a nonresponse mechanism exists. This implies that the meaning of the marginal reliability is different depending on the existence of a nonresponse mechanism. The marginal reliability is the mean reliability averaged over the distribution of the latent variable ξ and the distribution of missingness given by $\boldsymbol{D}$. This can be illustrated considering Data Example A. As Table 3.3 shows, the difference between the marginal reliability coefficients of the complete and the incomplete data of Data Example A is more than 0.15. However, the test information functions were only slightly different (see Figure 3.20). So, the marginal reliability depends not only on the test and the distribution of ξ but also on the nonresponse mechanism of the considered population. Why are these considerations important? Consider the case where a single population is studied. Two representative samples A and B are drawn. Sample A is assessed by a high-stakes assessment, data in Sample B were obtained by means of a low-stakes assessment. As expected, the proportion of missing data in Sample A is much lower than in Sample B. In this case the marginal

reliability estimates will be considerably different even if person and item parameters can be estimated unbiasedly in both samples. In this example the motivation to complete the test affects the marginal reliability, while the test information function implied by item parameters remains unaffected. Insofar, the marginal reliability is no longer a measure of the mean accuracy of person parameter estimation by the test, but the mean accuracy of person parameter estimation due to the test and the missing data mechanism.

This is also true for the marginal EAP reliability that was shown to be the variance ratio $Rel(\hat{\xi}_{EAP}) = Var(\hat{\xi}_{EAP})/Var(\xi)$ (Adams, 2005; Mislevy et al., 1992). Due to missing data, the variance $Var(\hat{\xi}_{EAP})$ decreases due to an increased shrinkage effect (see Table 3.3). This results in lower marginal reliabilities.

Figure 3.21 shows the average marginal reliabilities observed in the simulation study. A detailed analysis of the simulation results revealed that the sample size did not influence the marginal reliabilities under the simulated conditions. For that reason, each cell of Figure 3.21 gives the mean marginal reliabilitiy of 150 data sets simulated under three sample size conditions ($N = \{500, 1000, 2000\}$). The attenuation of marginal reliabilities caused by missing data is different for ML, WML, and EAP estimates. The correlation $Cor(\xi, \theta)$ is of minor importance. Even if the missing data mechanism is MCAR, the reliability decreases. The attenuation is mainly driven by the proportion of missing data and the number of variables Y_i in the measurement model. The marginal reliabilities of the EAP estimates are generally less attenuated, while the reliability of the WML estimates proofed to be mostly decreased by missing data.

3.4 Discussion

In this chapter the impact of missing data on sample-based estimates of item and person parameters were studied twofold - analytically and by means of simulation. Results of previous studies with real data suggested that IRT parameters might be pretty robust even if the nonresponse mechanism w.r.t. $\boldsymbol{Y}$ is NMAR (Culbertson, 2011, April; Pohl, Gräfe, & Hardt, 2011, September; Rose et al., 2010). Hence, it could be argued that ignoring missing data is admissible. Indeed, IRT parameter estimates seem to be less sensitive to missing data compared to CTT-based item and person parameter estimates. However, it could be demonstrated that increasing proportions of nonignorable missing data also result in biased IRT item and person parameter estimates. This highlights the need for appropriate approaches to handle item nonresponses. In the following sections the findings are briefly summarized.

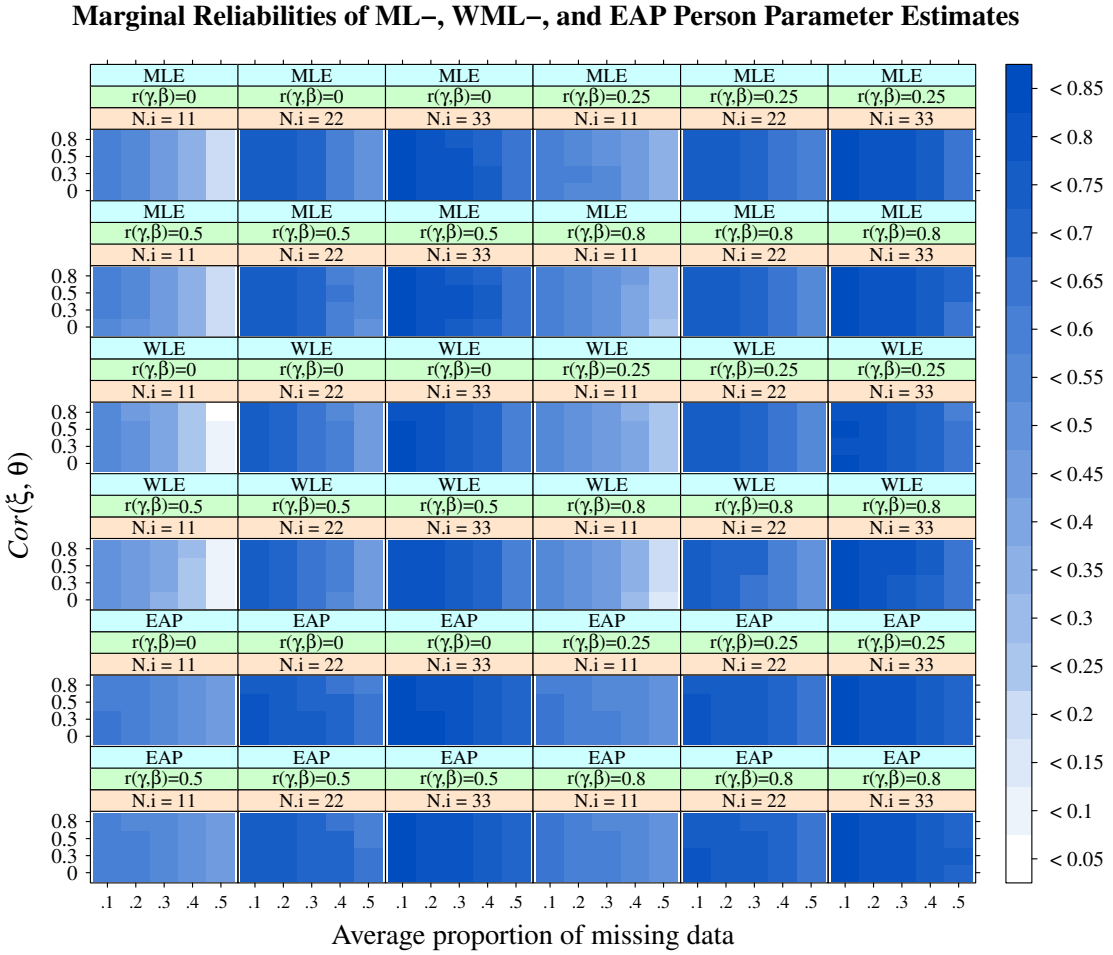

Figure 3.21: Marginal reliabilities of ML-, WML-, and EAP- person parameter estimates (simulation study).

3.4.1 Analytical Findings

Unfortunately, the use of analytical methods to study the impact of missing data is limited. Primarily, CTT-based item and person parameters can be studied analytically. Here the expected values $E(Y_i)$ as measures of item difficulties and the sum score S and the proportion correct score $P+$ as person parameter estimates were considered.

Sum score The **sum score** S or functions $f(S)$ are commonly used in CTT as person parameter estimates. It could be shown that S of a completely observed response pattern is a different random variable than S_{Miss}, the sum score in presence of missing data. The latter can formally be written as the sum of the I product variables $Y_i \cdot D_i$, implying an

implicit missing data handling. Item nonresponses are scored as $Y_i = 0$. Generally, $Y_i \cdot D_i$ and Y_i are different variables with different distributions if $P(Y_i = 1 \mid D_i = 0) > 0$. Hence, if there is a probability greater than zero to solve an omitted item, then the sum score is generally negatively biased under any missing data mechanism. Particularly worrying is that the implicit coding of item nonresponses as wrong responses leads to a confusion of two pieces of information: (a) the performance on the test items expressed by the items Y_i, and (b) the willingness or ability to respond to item i indicated by D_i. Hence, S_{Miss} in presence of missing data has a different meaning compared to S in absence of missing data. These analytical findings have implications with respect to ad hoc methods used in IRT models to handle item nonresponses. The coding of missing data as wrong responses, called Incorrect Answer Substitution (IAS), is a well- known and still widely used ad hoc method to handle item nonresponses. As in the case of S_{Miss}, the items in the measurement model Y_i are replaced by $Y_i \cdot D_i$. This potentially changes the meaning of the latent variable constructed in an IRT measurement model. These findings highlight that missing data and their improper handling are a threat of validity of test results. The consequences of IAS in IRT models will be examined in more detail in Section 4.3.1.

Proportion correct score The proportion correct score P^+ can be regarded as an individually standardized sum score. The sum score S_{Miss} is divided by the number of completed items. By using a simulated data example, it could be demonstrated that the bias is different compared to the sum score. Whereas the sum score can only be negatively biased, the proportion correct score can be negatively or positively biased. However, here it was argued that in most real applications P^+ is expected to be positively biased. The reason is that empirical findings support the hypothesis the intentionally omitted items are not arbitrarily skipped. Typically more difficult items are omitted with higher probabilities than easier items. Persons who tend to respond only to easier items will tend to have higher a proportion of correct scores than equally proficient persons who answer difficult items as well. This would lead to a positive bias of P^+. In testings with time limits, the bias of P^+ due to not-reached items depends on the item difficulties of the last items. Especially when extremely difficult or easy items are placed at the end of the test, P^+ will be positively or negatively biased. For example, when tests are applied with items that are ordered due to their difficulties and the time of the test is limited, the proportion correct score is not an appropriate test score. In summary, the proportion correct score accounts for item nonresponses but not sufficiently, since differences between responded and omitted items are not considered.

Item means The item mean of item i computed by observed item responses to item i are estimates of $E(Y_i \mid D_i = 1)$ instead of $E(Y_i)$. If $Y_i \not\perp D_i$ and measurement invariance w.r.t. Y_i given D_i hold, then stochastic dependence $D_i \not\perp \xi$ and systematically biased item means are implied. The reason is that the expected values $E(Y_i \mid D_i = 1)$ are computed by the integration over the conditional distribution $g(\xi \mid D_i = 1)$. The conditional distributions $g(\xi \mid D_i = 1)$ and $g(\xi \mid D_j = 1)$ $(i \neq j)$ can be different depending on the missing data mechanism with respect to the single items Y_i and Y_j respectively. Thus, each item of a single test is potentially answered by a different population when the missing data mechanism is MAR or NMAR. Since the expected values are population specific measures of item difficulty, the sample-based item means are measures that refer to unknown populations with respect to the distribution of the latent variable.

3.4.2 Simulation Study

Since IRT parameter estimates and their bias es can hardly be studied analytically, a simulation study was used. The estimation equations used to obtain item and person parameter estimates were considered. The interdpendence of unbiased item and person parameter estimation was shown. Although the biasedness of IRT parameter estimates is difficult to study theoretically, the analytical findings with respect to the bias of CTT-based item and person parameter estimates suggest that IRT-based parameters are potentially affected by item nonresponses as well.

IRT item difficulties When the missing data mechanism w.r.t. $\boldsymbol{Y}$ is NMAR, each item is potentially answered by a different population of test takers who differ with respect to their distribution of the latent ability ξ. It was expected to find negatively biased item difficulties, if more difficult items are omitted with higher probabilities and the tendency to omit items is positively correlated with the latent ability ξ. This expectation rests upon the finding that more difficult items are answered by persons with, on average, higher ability levels, while easier items are answered by persons with lower ability levels. The results of the simulation study confirmed this hypothesis. The negative bias of the estimates $\hat{\beta}_i$ is mainly driven by the correlation between the latent response propensity and the latent ability, and the overall proportion of missing data. These two factors explained 38 % of the variance the mean bias. In contrast, no bias was found when the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MCAR even when the overall proportion of missing data was 50 %.

IRT item discriminations The pattern of biases found in item discriminations is quite different from that of the item difficulties. The most important factors determining the bias of $\hat{\alpha}_i$ were the sample size and the number of items in the measurement model. Especially when the sample size was small ($N = 500$), the item discriminations were on average positively biased. The correlation $Cor(\xi, \theta)$ and the overall proportion of missing data had much less impact on discrimination parameter estimates than on item difficulty and person parameter estimates. Exceptionally, when the correlation between the latent ability and the latent response propensity was high ($Cor(\xi, \theta) = 0.8$), a small but consistent negative bias of $\hat{\alpha}_i$ occurred even in large sample sizes $N = 2000$.

IRT person parameter estimates With respect to IRT-based person parameter estimates no direct hypothesis could be derived from the analytical considerations of CTT-based person parameter estimates S and P^+. In unidimensional Rasch- and Birnbaum models item difficulties are locations on the same scale as the latent variable ξ. Hence, the bias of item and person parameter estimates are potentially correlated. Therefore, negative bias of the estimated item difficulties may induce a negative bias in person parameters. This seemed to be likely especially if MML estimation is applied, because the estimated item parameter estimates are taken as fixed values for the estimation of person parameters. In fact, ML and Warm's WML estimates turned out to be negatively biased in the simulation study. The correlation of the mean biases between item and person parameter estimates was $r = 0.815$ for ML estimates, $r = 0.846$ for WML estimates, and $r = 0.604$ for EAP estimates. Accordingly, the pattern of bias across the conditions used in the simulation study is very similar between item difficulties and ML and WML estimates. The correlation of the latent response propensity and the latent ability, and the overall proportion of missing data were found to be the most important factors of the bias. Both explained 36 % (ML estimates) or 40 % (WML estimates) of the variance of the mean bias. The stronger the correlation is and the higher the proportion of missing data is, the more negative the bias of ML and WML estimates is. Since the bias of ML and WML estimates is nearly uncorrelated with the proportion of item nonresponses, the bias results mostly from biased item parameter estimates. Surprisingly, on average the EAP estimates were unbiased in the conditions investigated in the simulation study. However, the bias of the EAPs is negatively correlated with the latent ability intended to be estimated. This correlation reflects the shrinkage effect, which is characteristic for Bayesian estimates. However, with an increasing proportion of missing data the shrinkage effect is intensified, resulting in a considerable variance reduction in the EAP estimates and

potentially unfair test results.

Shortcomings of the simulation study As in each simulation study, the generalizability of the results is restricted to the conditions under study. Here, tests with a small to medium number of items and small to medium sample sizes were considered. In large scale assessments the sample sizes are typically much larger. In high-stakes testings instruments with more than thirty items are regularly used. The results cannot be generalized to such applications. Furthermore, the nonignorability of missing data in the simulation study was generated by using a latent response propensity that was correlated with the latent ability. This approach allowed easily to vary the degree of stochastic dependency between $\boldsymbol{Y}$ and $\boldsymbol{D}$. However, there might be alternative data generating models in real applications that do not involve a latent response propensity. Data Example A as well as the simulation study emulate foremost the case where item nonresponses result from omissions instead of not-reached items. The latter result if persons fail to complete all items in timed tests. This results in a typical monotone missing pattern. The data generating models used for Data Example A and the simulation study do not account for such item nonresponses. This additionally limits generalizability of the results of this simulation study.

In the interpretation of the results of the simulation study, the identification of the model needs to be taken into account. In all simulations the model was identified by fixing the scale of the latent variable ξ. Generally the expected values was fixed to $E(\xi) = 0$ and in the Birnbaum model the variance was constrained to be $Var(\xi) = 1$. Alternatively, the models could have been identified by fixing an arbitrary item difficulty or the mean of the item difficulties, and by fixing at least one of the item discriminations in the case of the 2PLM. The bias of parameter estimates will probably be different with these model specifications. The bias is potentially transferred to other parameter estimates such as those describing the distribution of the latent variables. Conclusively, the ML and WML person parameter estimates or EAPs could be biased differently. Therefore, which parameters are estimated with a bias and the extent of the bias can depend on the identification of the model.

Despite the lack of generalizability, the results highlight that item nonresponses should be taken seriously in real applications when IRT models are used. This is all the more important, the stronger the dependency between missingness ($\boldsymbol{D}$) and the measurement instrument ($\boldsymbol{Y}$) is, and the higher the proportion of missing data is. This underlines the importance of appropriate approaches to handle item nonresponses.

3.4.3 Item Nonresponses and Test Fairness

Although there is a lack of a unique and widely accepted definition of test fairness (Kunnan, 2004), all approaches agree that construct-irrelevant sources of item and test difficulty threatens comparability of test scores and therefore test fairness (Zieky, 2006). For example, the analysis of differential item functioning and differential test functioning (Shealy & Stout, 1993) aims to identify such sources. The study of the bias of the sum score and the proportion correct score suggests that test fairness is also affected by item nonresponses. Hence, missing data and the way to handle them are potentially a source of construct-irrelevant variance in test scores and person parameter estimates. The implicit coding of item nonresponses to $Y_I = 0$ when the sum score S is used is a kind of penalization of persons who tend to omit items or fail to reach the end of the test. If test takers differ with respect to their tendency to respond to items in the test, then they will differ in the expected sum score $E(S_{Miss} \mid U)$ even if they have the same value of the latent ability ξ. On the one hand this reflects the change in the meaning of the sum score in presence of missing data, on the other hand this can be seen as a lack of test fairness depending on the intended meaning of the resulting test scores.

Although quite differently affected, the proportion correct score P^+ proved to be most likely biased in most applications. A prerequisite of comparability of proportion correct scores between test takers is that they answered the same test. However, due to omission of items each test taker creates his or her own test. The most likely scenario was considered exemplarily, where persons with lower ability levels prefer to answer easier items while tending to skip more difficult items. In this case the mean test difficulty T_β is stochastically dependent on the latent ability and P^+ are not comparable across persons. This leads to a higher proportion correct scores for persons with item nonresponses compared to those that complete all items even if they have equal proficiency levels. Insofar, omitting o difficult items becomes an attractive and beneficial response alternative when the proportion correct score is used as a test score. Similarly, EAP-scores tend to shrink toward the mean. The shrinkage effect is stronger, the less data are available. Therefore, the shrinkage effect varies across test takers depending on the proportion of item nonresponses. Increasing correlations between the EAP-bias and the latent ability by increasing proportions of missing data were found. This implies that below-average test takers would profit from omissions of items while above-average persons would be penalized for item nonresponses when EAP scores are used. Persons with ability levels below the average will increasingly profit from the shrinkage effect with rising proportions of omitted or not-reached items.

ML and WML estimates do not suffer from the shrinkage effect. Furthermore, the bias of both person parameter estimates is nearly uncorrelated with proportion of missing data. The issue of test fairness is of minor importance when these person parameter estimates are used.

3.4.4 Reliability

Reliability was examined with focus on IRT person parameter estimates. The standard error function and the marginal reliability were considered.

Standard error function In absence of any missing data mechanism, the standard error function is a function of the latent variable whose functional form depends solely on the item parameters. In presence of missing data, standard errors depend additionally on the missing pattern $\boldsymbol{D}$. Strictly speaking, there exist as many standard error functions as missing data pattern minus one. Each missing pattern is associated with a different subset of items and, therefore, a different standard error function according to the corresponding selection of items. Hence, if the test information function and the standard error function are estimated based on item parameter estimates, the resulting functions only refer to persons with complete response vectors. Both functions will be consistently estimated if the item parameter estimates are unbiased. However, this item information and standard error function are not meaningful with respect to persons with item nonresponses.

Marginal Reliability It was shown that that meaning of the marginal reliability changes if a nonresponse mechanism exists w.r.t. $\boldsymbol{Y}$. If no missing data mechanism exists, the marginal reliability depends only on the items in the test and the distribution of the latent variable ξ. If a missing data mechanism exists, then the marginal reliability depends on not only on the distribution of ξ and the test items, but also on the distribution of $\boldsymbol{D}$. Accordingly, the interpretation of the marginal reliability is affected. Without missingness, the marginal reliability can be interpreted as the average reliability of the person parameter estimates with respect to a particular population with its specific distribution of the latent variable. Under any missing data mechanism as defined in Section 2.2, the marginal reliability is the average reliability of the person parameter estimates with respect to a particular distribution of the latent variable and given the particular distribution of missingness ($\boldsymbol{D}$). Therefore, the marginal reliability can be substantially different between low-stakes and high-stakes assessments even if the same test would be applied to same sample due to changes in the distribution of $\boldsymbol{D}$. In high-stakes assessments the

tendency to omit items is typically much lower. This reduces standard errors of person parameter estimates and, therefore, increases the marginal reliability although neither the distribution of the latent variable nor the item parameters have changed.

In summary The results of the bias analyses highlight that missing data affect different parameter estimates differently and sometimes in an unexpected way. Furthermore, item nonresponses are a construct-irrelevant source of variability in test scores implying that test fairness as well as validity are potentially threatened. Although pretty robust, IRT item and person parameter estimates were also found to be consistently biased if the non-response mechanism w.r.t. $\boldsymbol{Y}$ is NMAR. This underlines the requirement of appropriate approaches for item nonresponses.

4 Missing Data Methods in Educational and Psychological Testing

In the previous section the need for appropriate methods to handle item nonresponses was demonstrated. In this section different approaches to handle missing data in educational and psychological measurement will be studied. Most of these approaches are not distinctive to the field of measurement. Rather they refer to well-known and widely used classes of missing data handling methods, which are briefly introduced in the beginning. In application IRT parameters can be estimated using ML estimation or Bayesian estimation procedures. This work focuses on ML estimation, in particular MML estimation with and without missing data. To clarify the terminology used in the remainder, ML estimation will be reviewed in Section 4.2. Although often criticized, the treatment of item nonresponses as incorrect answers is still common practice of achievement tests. Alternatively, missing responses are regularly scored as partially correct. Both approaches are critically examined in light of modern missing data handling methods in Sections 4.3.1 and 4.3.2. More recently, it was proposed to consider missing responses as an additional response category. The applicability of this approach is examined considering the implicit assumptions of this approach (see Section 4.4). The major focus of this work lies on multidimensional IRT (MIRT) models for nonignorable item nonresponses which are scrutinized in Section 4.5. This is done with the focus on the explicit and implicit underlying assumptions in these models. Typically, alternative MIRT models for item nonresponses have been considered to be equivalent in the literature (e. g. Holman & Glas, 2005; Rose et al., 2010). In fact, however, they are not necessarily equivalent. The conditions that ensure that missing data models are equivalent will be outlined. Based on these considerations alternative models will be derived. Furthermore, the classes of IRT models for nonignorable item nonresponses will be extended. Less restrictive MIRT models are proposed and latent regression models (LRM) (see Section 4.5.4) and multiple group (MG) IRT models (see Section 4.5.5) are introduced as alternatives to MIRT models for missing data. Finally, it will be demonstrated that item nonresponses due to omissions cannot equally be treated as missing responses due to not-reached items in MIRT models. For this reason a

joint model for omitted and not-reached items is introduced in Section 4.5.6.

This work focuses mainly on models for nonignorable missing data. The reason is that well known approaches for ignorable item nonresponses have been developed. These will be briefly reviewed in Section 4.5.2 using the example of computerized adaptive testing (CAT) with and without a routing test. Although ignorable missing responses are of minor interest here, especially models for item nonresponses that are MAR given $\boldsymbol{Z}$ are worth considering due to their close relation to models for nonignorable missing data.

4.1 Introduction To Missing Data Methods

In this section a short review of existing methods to handle missing data will be given in order to integrate methods used for item nonresponses in educational and psychological measurements. Several classification schemes of missing data handling methods have been proposed in the literature (Allison, 2001; Little & Rubin, 2002; Lüdtke et al., 2007; McKnight et al., 2007; T. Raghunathan, 2004; Schafer & Graham, 2002), that are the basis for the taxonomy used here. However, the list of methods considered in this classification is not exhaustive. The considerations are confined to the most important approaches that are relevant in the discussion about handling item nonresponses in measurement.

Analysis based on complete and available cases Simply to ignore the missing data is still the most commonly used practice (McKnight et al., 2007). For instance, the so-called complete case analyses include all of the observations without missing data while discarding those observations with incomplete data. This is commonly referred to as listwise deletion. This approach is not necessarily wrong with respect to biasedness of parameter estimates. However, analyses of complete cases assume that the missing data mechanism is MCAR. The advantage is that the reduced data set can be analyzed by standard estimation procedures for complete data. However, the amount of missing data is actually increased by eliminating data of test takers with incomplete data. The waste of a tremendous amount of useful and proverbially expensive information is unacceptable. The problem of item-nonresponse is replaced by the problem of unit-nonresponse. Due to the reduced sample size less information is available for estimating model parameters. Thus, listwise deletion is not efficient and results in a loss of precision reflected by larger standard errors. Complete case analysis becomes critical when the excluded persons might systematically differ from the persons that remain in the analysis. In this case, missing data mechanism is MAR or NMAR and listwise deletion can lead to seriously

biased parameter estimates. In educational and psychological measurement the crucial questions is whether the probability of non-responses is related to the items Y_i. Formally, is there a stochastic relationship between Y_i and D_i? If so, the missing data mechanism is not MCAR and potentially biased item and person parameter estimates result from listwise deletion. The MCAR assumption is very strong and hardly tenable in most psychological and educational measurements if missing responses result from omitted or not reached items.

Furthermore, complete case analysis is simply not applicable to test designs with planned missing data that are commonly used in many large scale assessments. For example, in multi-matrix sampling designs a booklet with a selection of items is assigned to each test taker. Due to not-administered items there are no cases with complete data and the effective sample size using listwise deletion is zero.

Analysis based on available cases refers to pairwise deletion and is mostly discussed in the context of linear regression analysis, factor analysis, and SEM where the model parameters can be estimated based on summary statistics such as means, variances, and covariances (Allison, 2001). Pairwise deletion means to use all observed data points in the computation of these summary statistics. This can be regarded as listwise deletion in the computation for each mean, variance, and covariance, separately. As a result, each estimated summary statistic is based on a different subsample that potentially differs systematically if the missing data mechanism is not MCAR. Furthermore, the number of observations used to calculate the summary statistics can vary considerably. Accordingly, the estimation of test statistics and standard errors is challenging and biased in most available software packages regardless of the missing data mechanism. Unfortunately, covariance matrices obtained by pairwise deletion are frequently not positive definite even if the missing data mechanism is MCAR.

In IRT models, both, complete and available case methods, are of minor importance. Commonly used ML estimation includes all observed item responses and is not based on bivariate (tetrachoric) correlations. However, SEM for dichotomous and ordered categorical data (Muthén, 1984) is an alternative to estimate item and person parameters of one- and two parameter probit models (Kamata & Bauer, 2008; Takane & de Leeuw, 1987). Model estimation with missing data rests upon uni- and bivariate frequency tables and estimated tetra- and polychoric correlation matrices. In this approach, thresholds, polychoric correlations, and probit regressions need to be fitted in the beginning of the estimation process. As in traditional SEM, pairwise deletion is still commonly used since an equivalent to FIML for those models is currently not available (Asparouhov & Muthén,

2010).

Weighting procedures Different weighting procedures can be distinguished. The most common strategies rest upon weighting cases with complete data to adjust for the selection of observations due to the nonresponse mechanism. Following Little and Rubin (2002), systematic missing data can be regarded as a selection problem, so that particular subpopulations are underrepresented in the sample. This imbalance is removed by giving observations from underrepresented populations more weight in the estimation process. Weighting procedures are directly related to propensity score analysis conducted in other fields (Guo & Fraser, 2009). Actually, weighting procedures are a modification of complete case analyses (Little & Rubin, 2002). That is, in multivariate analyses the cases with missing data are excluded. The remaining complete cases are appropriately weighted. A popular method is inverse probability weighting (IPW; Kim & Kim, 2007; Little & Rubin, 2002; T. Raghunathan, 2004; Wooldridge, 2007), where the inverse response propensities $P(D = 1 \mid U = u)^{-1}$ are used as weights[1]. In real applications $P(D = 1 \mid U = u)$ is typically unknown. However, given the person variable U is conditionally stochastic independent from D given the potentially multidimensional covariate $\boldsymbol{Z}$, the response propensities are $P(D = 1 \mid U, \boldsymbol{Z}) = P(D = 1 \mid \boldsymbol{Z})$. The weights $P(D = 1 \mid \boldsymbol{Z} = z)^{-1}$ may be known or can be estimated for each case given the covariate $\boldsymbol{Z}$ using, for example, logistic regression models. Note that conditional stochastic independence $U \perp \boldsymbol{D} \mid \boldsymbol{Z}$ implies that the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$. In fact, most commonly used weighting procedures require that the missing data mechanism is ignorable. Although point estimators are simple to compute, the computation of correct standard errors in weighted estimation procedures is sometimes difficult. This is one reason why weighting procedures are only recommended, especially in large samples.

In most common weighting approaches, there is one weight assigned to each observational unit with complete data. This is appropriate in order to adjust for sample selection biases due to unit nonresponses. However, this may fail to correct for item nonresponses. In Section 2.3 it was demonstrated that each single item can be answered by a different subsample that refers to a different subpopulation in terms of the distribution of the latent ability variable. Furthermore, the item response propensities $P(D_i = 1 \mid U = u)$ can differ within a single person. How does one assign a single weight to each test taker in a joint measurement model of all items? Actually, each individual response needs to be

[1]The subscript i of the response indicator has been omitted, since each case has a single response propensity that applies to all considered variables.

weighted. Using IPW the weights are given by $P(D_i = 1 \mid U = u)^{-1}$. An I-dimensional vector of weights $\{P(D_1 = 1 \mid U = u)^{-1}, \ldots, P(D_I = 1 \mid U = u)^{-1}\}$ result for each test taker. Most statistical software, however, allow only for a single weight per observational unit. Hence, weighting procedures are hardly applicable in multivariate analyses with item nonresponses and have been rarely addressed in the literature (e. g. Moustaki & Knott, 2000).

Imputation based methods Imputation based methods have become very popular in the recent years (Graham, 2009; Rubin, 1996; Schafer & Graham, 2002). Especially multiple imputation (MI) has proved to be an appropriate approach to account for missing data. The underlying idea of all currently used imputation methods is to replace missing responses by more or less plausible values. The completed filled-in data sets can be analyzed with standard methods for complete data. Hence, MI is a stepwise procedure consisting of (a) the augmentation of incomplete data sets, (b) the analyses of filled-in data sets, and (c) the combination of the results from the multiply imputed data to obtained point estimates and correct standard errors. Similarly, test statistics, such as the likelihood-ratios and p-values, can be combined (Schafer, 1997). The last step is dropped in single imputation methods. However, each imputation method starts with the modeling task (Little & Rubin, 2002; Rubin, 1987) that requires the specification of an imputation model. The imputation model specifies how to impute missing values based on observed data $\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}$. Unbiased sample based inference using imputation methods rests upon the correct specification of the imputation model. For example, the imputation model of MI with sequential regressions or chained equation (T. Raghunathan et al., 2001; Van Buuren, 2007) consists of linear or nonlinear regressions of each variable with missing data on the remaining variables in the data set and distributional assumptions with respect to the residuals of these regressions. If the regressions are correctly specified and the distributional assumptions hold true, the filled-in data sets can be seen as realizations $\boldsymbol{y}$ of $\boldsymbol{Y}$ with the distribution $g(\boldsymbol{Y})$. As shown in Section 2.2 (pp. 24 - 26), the latter can be written as the joint distribution $g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs})$ that can be factored into $g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \mid \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_{mis}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs})$. The first factor is the predictive distribution (e. g. Little & Rubin, 2002; Schafer, 1997). $\boldsymbol{\iota}_{mis}$ is the vector of regression coefficients and residual variances and covariances. Using MI, imputed values are random draws from the predictive distribution. Apart from MI, many imputation methods exist that differ with respect to the complexity of the imputation model and the respective assumptions. For example, there exist several naive approaches to handle item nonresponses such as

item mean substitution and person mean substitution (Huisman, 2000). In these cases, missing responses to item i are replaced by the item mean $\bar{y}_i$ or the proportion correct score P^+ of the completed items. Hence, the imputation model is simply an assignment rule. The frequently criticized but still often used incorrect answer substitution (IAS) is also a naive imputation method preferentially applied in achievement tests. The missing responses are scored as incorrect answers ($Y_i = 0$). Obviously, the imputed data sets can be very different depending on the used imputation method. Accordingly, the parameter estimates and their statistics will differ as well. The variance of the results suggests that the choice of the imputation method is essential. Given the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR, MI has proved to be an excellent method to handle missing data. With the introduction of sequential or chained regressions (T. Raghunathan et al., 2001; Van Buuren, 2007, 2010) MI has also become applicable in measurement models with binary and categorical manifest variables. Recent simulation studies proved MI to be useful for item nonresponses even if the proportion of missing data exceeds the proportion of the observed data considerably (Van Buuren, 2010). Although Rubin (1987) discussed MI for the case of nonignorable missing data as well[2], most of the currently implemented MI algorithms requires that the MAR assumptions hold true. In real applications omitted and not-reached items are typically related to test performance and, therefore, to persons' proficiency levels (Culbertson, 2011, April; Rose et al., 2010). Hence, the missing data mechanism is most likely nonignorable and MI is not appropriate. For that reason MI is not further considered in this work.

However, naive imputation methods are still commonly used even in large prestigious educational assessments such as PISA (Culbertson, 2011, April; Rose et al., 2010). The simplicity of such methods and their plausibility are tempting. For that reason, IAS and scoring missing responses as partially correct (PCS) are examined with respect to the implicit imputation model and the respective assumptions in Sections 4.3.1 and 4.3.2. The questions of whether and when these methods are appropriate to handle item nonresponses will be answered.

Model-based methods Model-based approaches estimate parameters of the target model directly from the incomplete data set. Missing data are directly taken into account in the model estimation. Compared to imputation methods model-based approaches are single

[2] If the nonresponse mechanism is NMAR, then the predictive distribution (see page 25) includes the response indicator variables D_i. Hence, $g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}_{mis})$. However, the estimation of the parameters $\boldsymbol{\iota}_{mis}$ of the imputation model is difficult, limiting the application of MI for nonignorable missing data.

step procedures. Nevertheless, many model-based approaches and imputation methods are closely related. The basic idea is quite simple. Following Little and Rubin (2002), missing data need not to be replaced by random draws from the predictive distribution. Instead, conditional expectations of missing values given observed values can be substituted for item nonresponses directly into the estimation equations. The resulting ML estimator comprises the estimation of the parameters of the target model and the parameters that relate observable and missing variables. The latter are equivalent to the parameters $\boldsymbol{\iota}_{mis}$ in an imputation model. Different ML estimators as well as Bayesian models have been developed to account for item nonresponses. Well known examples are the full information maximum likelihood (FIML) estimation (Arbuckle, 1996; Enders, 2001b) and the expectation-maximization (EM) algorithm (Dempster et al., 1977). Model based approaches have been developed for both ignorable and nonignorable missing data. For that reason they are of major interest here in this work.

Strictly speaking, sample based inference in presence of missing data is conditional given the observed missing pattern $\boldsymbol{D} = \boldsymbol{d}$. Since $\boldsymbol{D}$ is itself a random variable, sample based inference needs to be based on a joint model of $(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{D})$. Hence, the response indicator variables need to be modeled jointly with $\boldsymbol{Y}$ and $\boldsymbol{Z}$ as the variables of the target model. In fact, in Section 4.5.1 it will be shown in detail that the likelihood function that accounts for missingness is proportional to the joint distribution $g(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{D})$. Unfortunately, the specification and identification of models including $\boldsymbol{D}$ is quite difficult in many applications. Additionally, the model that reflects researchers' theory does not typically involve the response indicator variable $\boldsymbol{D}$. Hence, the model becomes pretty complex. Therefore, the statistical literature has extensively discussed the requirements that are needed to skip $\boldsymbol{D}$ from the parameter estimation of the target model. In his seminal paper, Rubin (1976) examined the weakest conditions that allow for ignoring $\boldsymbol{D}$ without affecting sample based inference. He proved that $\boldsymbol{D}$ needs not to be included in ML and Bayesian estimation if the missing data mechanism is ignorable (MCAR or MAR). If the nonresponse mechanism is NMAR, then the missing data are nonignorable, meaning that $\boldsymbol{D}$ cannot be ignored in ML and Bayesian parameter estimation. IRT models can be estimated by ML or Bayesian methods. The latter will not be considered here. ML estimation of IRT models with missing data will be examined in detail in Section 4.5.1. In general, ML estimation is briefly reviewed and summarized in the subsequent section.

Selection models (SLM) (Heckman, 1976, 1979; Little, 2008; Winship & Mare, 1992) and pattern mixture models (PMM; Little, 1993, 2008) are two classes of model based approaches for nonignorable missing data. Both approaches rest upon a joint model of

$\boldsymbol{Y}$ and the respective response indicator vector $\boldsymbol{D}$. In this work it will be shown that IRT models for nonignorable item nonresponses can be derived from SLMs or PMMs under certain assumptions. Such models for missing responses in IRT measurement models will be examined and further developed in Section 4.5.

4.2 Maximum Likelihood Estimation Theory

In the study of the bias of item and person parameter estimates (see Chapter 3) the terms maximum likelihood estimation and likelihood function have already been used. In this section, ML estimation is briefly reviewed in more detail since model based approaches considered in the remainder of this work are based on ML estimation. First, ML estimation with complete data is introduced. ML estimation in presence of different missing data mechanisms will be examined in Section 4.5.1.

Let there be a I-dimensional random variable $\boldsymbol{Y}$. N denotes the sample size. That is, the number of repetitions of the single unit trial as described in Section 2.2 (see Equations 2.7 and 2.8). The data matrix $\boldsymbol{y}$ is then a realization of an $N \times I$-dimensional random matrix $\boldsymbol{Y}$. Each row $\boldsymbol{Y}_n$ $(n = 1, \ldots, N)$ of $\boldsymbol{Y}$ represents a randomly drawn observational unit. For example, in a psychological test that is the response vector $\boldsymbol{Y}_n = Y_{n1}, \ldots, Y_{nI}$ of the n-th test taker. In the remainder, it is assumed that stochastic independence $\boldsymbol{Y}_n \perp \boldsymbol{Y}_m$ $(\forall\, n \neq m \in 1 \ldots N)$ holds. That is, the single unit trials are conducted independently. Let there be a parametric model with the parameter vector $\boldsymbol{\iota}$. The ML estimation of $\boldsymbol{\iota}$ rests upon the likelihood function $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$ that can be derived from the conditional probability function $g(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\iota})$. However, the function $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$ is not required to be a probability function (Enders, 2005; Held, 2008). It is sufficient that $\mathcal{L}(\boldsymbol{y};\, \boldsymbol{\iota})$ is proportional to $g(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{\iota})$. Thus, the likelihood function or simply the likelihood of $\boldsymbol{Y} = \boldsymbol{y}$ is proportional to the joint distribution of the N response vectors $g(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_N = \boldsymbol{y}_N; \boldsymbol{\iota})$. If the rows of $\boldsymbol{Y}$ are stochastically independent, then

$$\mathcal{L}(\boldsymbol{y};\, \boldsymbol{\iota}) \;\propto\; \prod_{n=1}^{N} g(\boldsymbol{Y}_n = \boldsymbol{y}_n; \boldsymbol{\iota}). \tag{4.1}$$

Let $\hat{\boldsymbol{\iota}}$ be an estimator of $\boldsymbol{\iota}$. The defined set of values that $\hat{\boldsymbol{\iota}}$ can take on is called the parameter space $\Omega_{\boldsymbol{\iota}}$. The ML estimator $\hat{\boldsymbol{\iota}}_{ML}$ of $\boldsymbol{\iota}$ is defined as the value of the parameter space $\Omega_{\boldsymbol{\iota}}$ that maximizes the joint probability density function and, therefore, the likelihood

function $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$:

$$\hat{\boldsymbol{\iota}}_{ML} = \arg\max_{\boldsymbol{\iota} \in \Omega_{\boldsymbol{\iota}}} \mathcal{L}(\boldsymbol{y};\ \boldsymbol{\iota}) \tag{4.2}$$

Since $\hat{\boldsymbol{\iota}}_{ML}$ is the maximizer of $L(\boldsymbol{y}; \boldsymbol{\iota})$, the estimation problem is equivalent to finding the roots of the first derivative $\mathcal{L}'(\boldsymbol{y}; \boldsymbol{\iota}) = \frac{\partial}{\partial \boldsymbol{\iota}} \mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$ with respect to $\boldsymbol{\iota}$. Typically, the natural logarithm of the likelihood $\ell(\boldsymbol{y}; \boldsymbol{\iota}) = \log[\mathcal{L}(\boldsymbol{y};\ \boldsymbol{\iota})]$ is maximized instead of the likelihood[3]. This is equivalent since the logarithm is a monotone transformation and the values $\boldsymbol{\iota} \in \Omega_{\boldsymbol{\iota}}$ that maximizes $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$ and $\ell(\boldsymbol{y}; \boldsymbol{\iota})$ are identical. Thus, parameter estimates are obtained by setting the first derivative $\ell'(\boldsymbol{y}; \boldsymbol{\iota}) = \frac{\partial}{\partial \boldsymbol{\iota}} \ell(\boldsymbol{y}; \boldsymbol{\iota})$ equal to zero and solving for $\boldsymbol{\iota}$. In multi-parameter estimation problems, $\ell'(\boldsymbol{y}; \boldsymbol{\iota})$ is a vector of the partial derivatives of $\ell(\boldsymbol{y}; \boldsymbol{\iota})$ with respect to the single elements of $\boldsymbol{\iota} = \iota_1, \ldots, \iota_M$.

$$\ell'(\boldsymbol{y}; \boldsymbol{\iota}) = \frac{\partial \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \boldsymbol{\iota}} = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_1} \\ \frac{\partial \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_2} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_M} \end{pmatrix} \tag{4.3}$$

$\ell'(\boldsymbol{y}; \boldsymbol{\iota})$ is also called the gradient or the score vector. The second derivative $\ell''(\boldsymbol{y}; \boldsymbol{\iota})$ of the log-likelihood is the $M \times M$ Hessian matrix.

$$\ell''(\boldsymbol{y}; \boldsymbol{\iota}) = \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \boldsymbol{\iota}^2} = \begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_1^2} & \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_1 \partial \iota_2} & \cdots & \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_1 \partial \iota_M} \\ \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_2 \partial \iota_1} & \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_2^2} & \cdots & \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_2 \partial \iota_M} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_M \partial \iota_1} & \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_M \partial \iota_2} & \cdots & \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\iota})}{\partial \iota_M^2} \end{pmatrix} \tag{4.4}$$

The negative of the Hessian matrix is the observed information matrix $I(\boldsymbol{\iota})$ (Efron & Hinkley, 1978; Held, 2008). Inverting $I(\boldsymbol{\iota})$ gives an estimator of the variance-covariance matrix $ACOV(\hat{\boldsymbol{\iota}}_{ML})$ of the estimator $\hat{\boldsymbol{\iota}}_{ML}$. In general, the ML estimator is consistent and, therefore, asymptotically unbiased, asymptotically efficient, and asymptotically normal, so that $\sqrt{N}(\hat{\boldsymbol{\iota}}_{ML} - \boldsymbol{\iota}) \rightarrow N(\mathbf{0}, I(\boldsymbol{\iota})^{-1})$ (e. g. Green, 2012). This implies $\hat{\boldsymbol{\iota}}_{ML} \rightarrow N\left(\boldsymbol{\iota}, I(\boldsymbol{\iota})^{-1}\right)$ for large samples (e. g. Held, 2008). The standard errors of the estimates in $\hat{\boldsymbol{\iota}}_{ML}$ are obtained by the square root of the diagonal elements of $ACOV(\hat{\boldsymbol{\iota}}_{ML})$.

So far, ML estimation theory has been introduced for the case of completely observed

[3] In application, the value of $\mathcal{L}(\boldsymbol{Y}; \boldsymbol{\iota})$ becomes rapidly tiny potentially causing computational problems. Additionally, the log-tranformed likelihood can be easier handled mathematically.

data. This is sufficient to examine data augmentation methods such as incorrect-answer-substitution and partially-correct-scoring of missing data as well as the use of the nominal response model for missing responses. These approaches have in common that filled-in data sets are used for parameter estimation. Hence, all missing values are replaced or recoded and ML estimation methods for complete data are used. Consequently, ML estimation with missing data is required for model-based approaches and is considered in detail in Section 4.5.1. The suitability of these three methods for item nonresponses will be critically studied next.

4.3 Data Augmentation Methods Used in IRT Models

Especially in educational testings there is strong evidence that item-nonresponses and the latent ability of interest are stochastically dependent. It was repeatedly found that the proportion of missing data decreases with increasing ability levels (Culbertson, 2011, April; Rose et al., 2010). This is a typical finding especially in low-stakes assessments. For instance, in the PISA 2006 data a substantial correlation of $r = 0.33$ was found between the proportion correct score and the proportion of answered items (Rose et al., 2010). The higher probability of missing data in persons with lower test scores seems to justify the recoding of missing responses to incorrect responses ($Y_i = 0$). In achievement testings the method is also called incorrect answer substitution (IAS) (Huisman, 2000). Despite criticism of this approach almost 30 years ago by Lord (1974), among others, IAS is still widespread in large scale assessments as in PISA, (Rose et al., 2010). Obviously, IAS has not lost any of its attractiveness, notwithstanding the persistent criticism against this practice (e. g., Lord, 1974, 1983a; Ludlow & O'Leary, 1999; Rose et al., 2010). Apart from the plausibility at first sight, the easy applicability of IAS might be responsible for its wide use. Furthermore, some IRT programs might be tempting for applied researchers to use IAS. For instance, in BILOG 3 (Zimowski, Muraki, Mislevy, & Bock, 1996) the user can only choose between two alternatives to treat omitted responses in the parameter estimation stage: (a) treating missing responses as wrong responses, or (b) as partially correct. Applied researchers may unintentionally suggest that these two options are the best practice to handle item nonresponses. Advocates of IAS often argue that it is not important to consider why test takers fail to give the correct answer. From this perspective it is irrelevant to distinguish between a wrong response and a nonresponse when the correct answer was not given by a test taker. This argumentation seems to be plausible at first glance but is potentially incompatible with a chosen measurement model that reflects

theoretical assumptions about the response process. Additionally, IAS is associated with implicit assumptions that may unlikely hold in application. In this work, IAS is considered to be an imputation method. As previously discussed, imputation based methods are appropriate if the imputation model is correctly specified and the underlying assumptions hold true. IAS will be studied from this point of view.

As an alternative to IAS, Lord (1974, 1983a) proposed to treat missing data as partially correct. The rationale of this method is that each test taker u has a positive probability $P(Y_i \mid D_i = 0, U = u)$ to solve an item even if no answer is observed. Partially correct scoring (PCS) of item nonresponses as an alternative to IAS is also studied as an imputation method, since missing responses are implicitly replaced by constants. This will be demonstrated in Section 4.3.2. PCS is also commonly used in large scale assessments as an alternative to IAS. It is implemented in some IRT software such as BILOG 3. Similarly, the simplicity and plausibility of PCS as well as its implementation in existing software is tempting for applied researchers. The underlying assumptions have rarely been made explicit. This will be done here. In the next two sections IAS and PCS will be scrutinized with respect to their assumptions, theoretical implications, and practical consequences. In order to demonstrate its performance, both approaches will be applied to Data Example A.

4.3.1 Incorrect Answer Substitution for Item Nonresponses

Following Huisman (2000), IAS is a naive or simple imputation method. A prerequisite of correct sample-based inference is the correct specification of the imputation model. That includes that the explicit and implicit assumptions of this model have to hold true in application. That the imputation model used in IAS is unlikely to be appropriate is already implied by the bias found in the sum score (see Section 3.1.1). Recall that the sum score implicitly recodes missing responses into incorrect responses. It was found that the sum score is only unbiased if the probability $P(Y_i = 1 \mid D_i = 0, U = u)$ to solve a missing item is equal to zero.

From a theoretical point of view it was shown that IAS means to replace the variables Y_i with new random variables $Y_i^* = Y_i \cdot D_i$ (see Equation 3.7). Both variables Y_i and Y_i^* have most likely different distributions and refer to different random experiments. Recall that if no missing data mechanism exists, then the random experiment is to draw u of U, administer a test consisting of the items $Y_1, \ldots, Y_I$, and observe the item responses. In contrast, the random experiment given the missing data are treated as wrong means to draw a unit u of U randomly, administer a test consisting of the items $Y_1, \ldots, Y_I$, observe

the item responses of answered items, and recode item nonresponses to $Y_i = 0$. Thus, Y_i^* is a function $f(Y_i, D_i)$ of item i and the respective response indicator. In this case, $f(Y_i, D_i)$ is an assignment rule given by Equation 3.7. However, when IAS is considered an imputation method it can be asked what the implicit assumptions are that need to hold true in order to ensure unbiased item and person parameter estimation. Furthermore, the theoretical implications of these assumptions can be examined.

Implicit assumptions underlying IAS and their implications What is the imputation model under IAS? In contrast to MI, the imputed values depend not on other manifest or latent variables. Missing values of each test taker are replaced by zeros regardless of other item responses or covariates. Hence, IAS rests upon a deterministic model. The imputed values depend only on the missing data indicators D_i. The assignment rule (see Equation 3.7) determines how to augment the incomplete data set. What are the implicit assumptions underlying IAS? It is assumed that $P(Y_i = 1 \,|\, D_i = 0) = 0$. This implies for each test taker u of U, so that $P(Y_i = 1 \,|\, U = u, D_i = 0) = 0$. In fact, in Section 3.1.1 it was shown that the sum score S_{Miss} in presence of missing data is unbiased if $P(Y_i = 1 \,|\, U = u, D_i = 0) = 0$ holds true (see Equation 3.30). Furthermore, in this case the equality $Y_i = Y_i^*$ is implied. Hence, although the measurement model under IAS consists of I regressions $P(Y_i^* \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})$ instead of $P(Y_i \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})$, the construction of the latent variable remains unaffected and item and person parameter estimates will be unbiased. Furthermore, multiple imputations are not required since each imputed data set is completely the same if $P(Y_i = 1 \,|\, U = u, D_i = 0) = 0$ holds true. Nevertheless, although this implicit assumption justifies the use of IAS, it has considerable implications and causes serious theoretical inconsistencies. Again, IAS assumes that $P(Y_i = 1 \,|\, D_i = 0) = 0$ implying that Y_i is a constant given $D_i = 0$. A constant is always stochastically independent from any other random variable. Consequently, $Y_i \perp U \,|\, D_I = 0$. If Y_i is a dichotomous item in a latent trait model with a latent variable $\boldsymbol{\xi} = f(U)$, then $P(Y_i = 1 \,|\, D_i = 0, \boldsymbol{\xi}) = P(Y_i = 1 \,|\, D_i = 0) = 0$. Thus, if the imputation model under IAS is correct, then the assumption of conditional stochastic independence $Y_i \perp \boldsymbol{\xi} \,|\, D_i = 0$ is implied. Hence, the probability to solve an omitted or not-reached item is zero regardless of the proficiency levels of the test takers! This implicit assumption is untenable in most realistic applications. Interestingly, IAS also assumes that $P(Y_i = 1 \,|\, D_i = 1, \xi) = P(Y_i = 1 \,|\, \xi)$. Hence, if an item response is observed, then the respective IRT model applies. Apparently, there is a strong interaction effect between D_i and $\boldsymbol{\xi}$ with respect to Y_i, implying measurement invariance with respect to D_i. Further interaction effects between D_i and all other random variables

are implied, which are stochastically dependent from Y_i given $D_i = 1$. In psychological and educational testings there are many covariates captured in $\boldsymbol{Z}$ that are stochastically dependent on the achievement on the test and the test items, respectively. Hence, $P(Y_i = 1 \mid \boldsymbol{Z}) \neq P(Y_i = 1)$. However, the assumption $P(Y_i = 1 \mid D_i = 0) = 0$ justifying IAS implies that $Y_i \perp \boldsymbol{Z} \mid D_i = 0$. Specifically, $P(Y_i = 1 \mid D_i = 0, \boldsymbol{Z}) = P(Y_i = 1 \mid D_i = 0) = 0$. In this case D_i would moderate the stochastic relation between $\boldsymbol{Z}$ and Y_i. This is also true for any other item $Y_{j \neq i}$. If Y_i and Y_j indicate the same latent variable, then they are correlated if the latent ability variables have a non-zero variance. That is, $Y_i \not\perp Y_j \mid D_i = 1$. The implicit assumptions of the IAS imputation model, however, imply $Y_i \perp Y_j \mid D_i = 0$.

Interestingly, the interactions between D_i and Y_j or D_i and $\boldsymbol{Z}$ with respect to Y_i implied by IAS assumptions have also implications with respect to the missing data mechanism of Y_i. The item Y_i and the response indicator D_i cannot be stochastically independent, neither unconditionally nor conditionally given Z or other observable items $\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{i}}$. Hence, if IAS assumptions hold true, then the missing data mechanism is not allowed to be MCAR or MAR unless $P(Y_i = 1 \mid D_i = 1, \boldsymbol{Z}) = P(Y_i = 1 \mid D_i = 1) = 0$ and $P(Y_i = 1 \mid D_i = 0) = P(Y_i = 1 \mid D_i = 1) = 0$. However, in this case the items would be constants ($Y_i = 0$).

To sum up, from the implicit assumption $P(Y_i = 1 \mid D_i = 0) = 0$ of IAS follows that the nonresponse mechanism cannot be MCAR or MAR. Even if the missing data mechanism is NMAR, IAS implies that the items are stochastically independent of any other variable given $D_i = 0$, including the latent ability. In contrast, if item i is observable ($D_i = 0$), then a stochastic dependency between Y_i and the persons' proficiency is assumed, which is described by the model equations $P(Y_i = 1 \mid \boldsymbol{\xi})$ of the measurement model. Hence, a very strong form of DIF given D_i is implied. Here the view is taken that none of these implications following from IAS are tenable in application. The theoretical inconsistencies make the use of IAS obsolete. Nevertheless, in the remainder of this section the consequences of using IAS with respect to item parameter estimation and the construction of the latent variable will be demonstrated for the case of a measurement model with a unidimensional latent variable ξ.

ML estimation under IAS As in any other imputation method, the filled-in data set is analyzed by standard methods. Here ML estimation is considered. Using IAS, ML estimation of model parameters is based on the variables Y_i^* instead of Y_i. Assuming local stochastic independence and that no missing data mechanism exists, the likelihood of a

unidimensional latent trait model with dichotomous variables Y_i can be written as

$$\mathcal{L}(\boldsymbol{y};\boldsymbol{\iota}) \propto \prod_{n=1}^{N}\prod_{i=1}^{I} P(Y_{ni}=1\,|\,\xi;\boldsymbol{\iota})^{y_{ni}} P(Y_{ni}=0\,|\,\xi;\boldsymbol{\iota})^{1-y_{ni}}. \tag{4.5}$$

The vector $\boldsymbol{\iota}$ contains the model parameters aimed to be estimated. That is, the item and person parameters. In presence of any missing data mechanism, however, the likelihood under IAS can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}^*;\boldsymbol{\iota}) \propto \prod_{n=1}^{N}\prod_{i=1}^{I}\Big\{&\Big[P(Y_{ni}=1\,|\,\xi;\boldsymbol{\iota})^{y_{ni}} P(Y_{ni}=0\,|\,\xi;\boldsymbol{\iota})^{1-y_{ni}}\Big]^{d_i}\\ &\Big[P(Y^*_{ni}=1\,|\,\xi;\boldsymbol{\iota})^{y^*_{ni}} P(Y^*_{ni}=0\,|\,\xi;\boldsymbol{\iota})^{1-y^*_{ni}}\Big]^{1-d_i}\Big\}.\end{aligned} \tag{4.6}$$

Using the derivations from above, the likelihood can be simplified. Since IAS assumes $P(Y_i=1\,|\,D_i=0)=0$ implying $Y_i \perp \xi\,|\,D_i=0$, Equation 4.6 can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}^*;\boldsymbol{\iota}) \propto \prod_{n=1}^{N}\prod_{i=1}^{I}\Big\{&\Big[P(Y_{ni}=1\,|\,\xi;\boldsymbol{\iota})^{y_{ni}} P(Y_{ni}=0\,|\,\xi;\boldsymbol{\iota})^{1-y_{ni}}\Big]^{d_i}\\ &\Big[P(Y^*_{ni}=0\,|\,D_i=0)^{1-y^*_{ni}}\Big]^{1-d_i}\Big\},\end{aligned} \tag{4.7}$$

with $P(Y^*_{ni}=0\,|\,D_i=0)=1$. In the first factor y_{ni} is used instead of y^*_{ni} since $y_i = y^*_i\,|\,D_i=1$. This likelihood is similar to a zero-inflated logistic regression mixture model with known classes indicated by D_i. Interestingly, the factor $[P(Y^*_{ni}=0\,|\,D_i=0)^{1-y^*_{ni}}]^{1-d_i}$ reduces to $P(Y^*_{ni}=0\,|\,D_i=0)^{1-d_i}$ and does not contain any estimand of the substantive model and could actually be skipped without affecting the remaining ML estimates. However, the theoretically implied likelihood given in Equation 4.7 is not used in real applications. Using any IRT program under IAS means that the likelihood given by Equation 4.5 is used where the variables Y_i are replaced by Y^*_i. In fact, the following function is maximized under IAS:

$$f(\boldsymbol{y}^*;\boldsymbol{\iota}) = \prod_{n=1}^{N}\prod_{i=1}^{I}\Big[P(Y_{ni}=1\,|\,\xi;\boldsymbol{\iota})^{y_{ni}} P(Y_{ni}=0\,|\,\xi;\boldsymbol{\iota})^{1-y_{ni}}\Big]^{d_i} P(Y_{ni}=0\,|\,\xi;\boldsymbol{\iota})^{1-d_i}, \tag{4.8}$$

This is neither the likelihood function with respect to $\boldsymbol{Y}^*=\boldsymbol{y}^*$ nor $\boldsymbol{Y}=\boldsymbol{y}$. Therefore, it is only denoted as an estimation function $f(\boldsymbol{y}^*;\boldsymbol{\iota})$. Note that the last factor that represents the missing responses is now the same probability of an incorrect answer *given* the latent ability as in the case of the observed item responses. Hence, the stochastic independence

$Y_i \perp \xi \,|\, D_i = 0$ implied by IAS is not taken into account in this likelihood function. It is not distinguished between zeros due to an incorrect answer and zeros due to item nonresponses. The derivations reveal that the implications from the underlying assumptions in the construction of Y_i^* are fundamentally ignored at the estimation stage. The wrong estimation equation is maximized resulting in potentially strange estimators.

On the basis of these considerations the item parameter estimates under IAS are unlikely to be equal to complete data. Furthermore, the construction of the latent variable is potentially affected by replacing Y_i with Y_i^*. The latent variable might not only differ numerically but also with respect to its meaning, thus threatening the validity of the test. In a first step the effects of IAS on item parameters will be examined followed by the effects with respect to the latent variable. Analytical derivations will be illustrated by empirical results from Data Example A.

Effects of IAS to item parameter estimates Recall that under IAS the regressions $P(Y_i = 1 \,|\, \xi; \mathfrak{l})$ are replaced by $P(Y_i^* = 1 \,|\, \xi; \mathfrak{l})$. Since only zeros are imputed for missing responses higher proportions of incorrect responses result that should mimic more difficult items. Therefore, if a one- or two-parameter model is identified by fixing the latent variable to $E(\xi) = 0$[4] it is expected that the item difficulties β_i will be overestimated in both models using IAS. Furthermore, it is expected that the item discriminations α_i in the two-parameter model will be biased depending on the missing data mechanism. Given that the missing data mechanism is non-ignorable with $Cor(\xi, \theta) > 0$, it is expected that item discriminations will be positively biased. In contrast, a negative bias is expected if the missing data mechanism w.r.t. Y_i is MCAR. At first the case of non-ignorable missing data is considered. In the following derivations it is assumed that the regressions $P(Y_i = 1 \,|\, \xi)$ and $P(D_i = 1 \,|\, \theta)$ are monotonically increasing functions of the latent variables with values between zero and one. Since the regressions can also be nonparametric, the parameter vector $\mathfrak{l}$ will be omitted. Given that $Cov(\xi, \theta) > 0$, then $Cov[\xi, P(D_i = 1 \,|\, \theta)] > 0$. This implies that the probability of an item nonresponses increases, the lower the latent ability ξ is. Since the response category $Y_i^* = 0$ results not only from incorrect answers to item i but also from nonresponses that are increasingly likely with lower values of ξ, the ICC approaches faster to zero with decreasing values of ξ. A positively biased item discrimination results. This can also be shown mathematically, using Y_i^*. As examined in Equation 3.7, the regression $P(Y_i^* = 1 \,|\, U)$ can be written

[4] In the two-parameter model the variance $Var(\xi)$ needs to be fixed as well for free estimation of all item parameters.

as $E(Y_i \cdot D_i \,|\, U)$. Given $Cov(\varepsilon_{Y_i}, \varepsilon_{D_i} \,|\, U) = 0$ (see Equations 3.12 and 3.15) that is,

$$P(Y_i^* = 1 \,|\, U) \quad = \quad P(Y_i = 1 \,|\, U) \cdot P(D_i = 1 \,|\, U). \tag{4.9}$$

Assuming a latent response propensity variable $\theta = f_1(U)$ and the latent ability $\xi = f_2(U)$ exist, the regression $P(Y_i^* = 1 \,|\, U)$ can be replaced by the regression $P(Y_i^* = 1 \,|\, \xi, \theta) = E(Y_i^* \,|\, \xi, \theta)$. The latter can be written as

$$\begin{aligned} E(Y_i^* \,|\, \xi, \theta) &= E(Y_i \cdot D_i \,|\, \xi, \theta) && (4.10) \\ &= Cov(\varepsilon_{Y_i}, \varepsilon_{D_i} \,|\, \xi, \theta) + E(Y_i \,|\, \xi, \theta) E(D_i \,|\, \xi, \theta) && (4.11) \end{aligned}$$

Assuming that $Cov(\varepsilon_{Y_i}, \varepsilon_{D_i} \,|\, \xi, \theta) = 0$ and conditional stochastic independencies $Y_i \perp \theta \,|\, \xi$ and $D_i \perp \xi \,|\, \theta$ hold true, it follows

$$\begin{aligned} E(Y_i^* \,|\, \xi, \theta) &= E(Y_i \,|\, \xi) E(D_i \,|\, \theta) && (4.12) \\ &= P(Y_i = 1 \,|\, \xi) P(D_i = 1 \,|\, \theta). && (4.13) \end{aligned}$$

Using IAS only the latent variable ξ is included in the model. Hence, the measurement model is constituted by the regressions $P(Y_i^* = 1 \,|\, \xi)$ that can be written as

$$\begin{aligned} P(Y_i^* = 1 \,|\, \xi) &= E[E(Y_i^* \,|\, \xi, \theta) \,|\, \xi] && (4.14) \\ &= E[E(Y_i \,|\, \xi) E(D_i \,|\, \theta) \,|\, \xi] \\ &= E(Y_i \,|\, \xi) \cdot E[E(D_i \,|\, \theta) \,|\, \xi]. \end{aligned}$$

From $D_i \perp \xi \,|\, \theta$ follows $E[E(D_i \,|\, \theta) \,|\, \xi] = E[E(D_i \,|\, \theta, \xi) \,|\, \xi] = E(D_i \,|\, \xi)$ implying that

$$\begin{aligned} P(Y_i^* = 1 \,|\, \xi) &= P(Y_i = 1 \,|\, \xi) \cdot E[P(D_i = 1 \,|\, \theta) \,|\, \xi] && (4.15) \\ &= P(Y_i = 1 \,|\, \xi) P(D_i = 1 \,|\, \xi). \end{aligned}$$

It can be seen that the regression $P(Y_i = 1 \,|\, \xi)$ is weighted by the regression of the response indicator D_i on ξ. The values of the regression $P(D_i = 1 \,|\, \xi)$ are probabilities ranging between zero and one by definition. Therefore, the difference $P(Y_i = 1 \,|\, \xi) - P(Y_i^* = 1 \,|\, \xi)$ is always ≤ 0. This difference can also be written as $P(Y_i = 1 \,|\, \xi) - P(Y_i = 1 \,|\, \xi) P(D_i = 1 \,|\, \xi)$. If $Cov[\xi, P(D_i = 1 \,|\, \theta)] > 0$ implied by $Cov(\xi, \theta) > 0$, then the ICC referring to $P(Y_i^* = 1 \,|\, \xi)$ is steeper because it approaches faster to zero with decreasing values of ξ. Taken together, it is expected that the ICCs will be shifted to the right due to an

overestimated β_i and are expected to be steeper because of the positively biased item discrimination estimates $\hat{\alpha}_i$.

Data Example A was used to confirm the expected biases of item parameter estimates. Exemplarily, two items Y_3 and Y_{28} are considered at first. Y_3 is a comparably easy item ($\beta_3 = -2.00$) which is little affected by missing data. The overall response rate is $\bar{D}_3 = 0.893$. Item Y_{28} is a pretty difficult item with $\beta_{28} = 1.750$. The overall response rate is much lower $\bar{D}_3 = 0.257$. Figure 4.1 shows the ICCs of $P(Y_i|\xi)$ and $P(Y_i^*|\xi)$ of both items. The results for these two items show the expected result pattern. The ICCs are right-shifted indicating positively biased item difficulty estimates. As Table 4.1 shows, $\hat{\beta}_3 = -1.114$ which is higher than the true value $\beta_3 = -2.00$. Similarly, $\hat{\beta}_{28} = 2.455$ which is higher than the true value $\beta_{28} = 1.74$. The item discrimination of item Y_3 is close to one ($\hat{\alpha}_3 = 1.029$). However, item 28 with many missing responses showed a considerably overestimated item discrimination of $\hat{\alpha}_{28} = 1.765$. Table 4.1 shows the item parameter

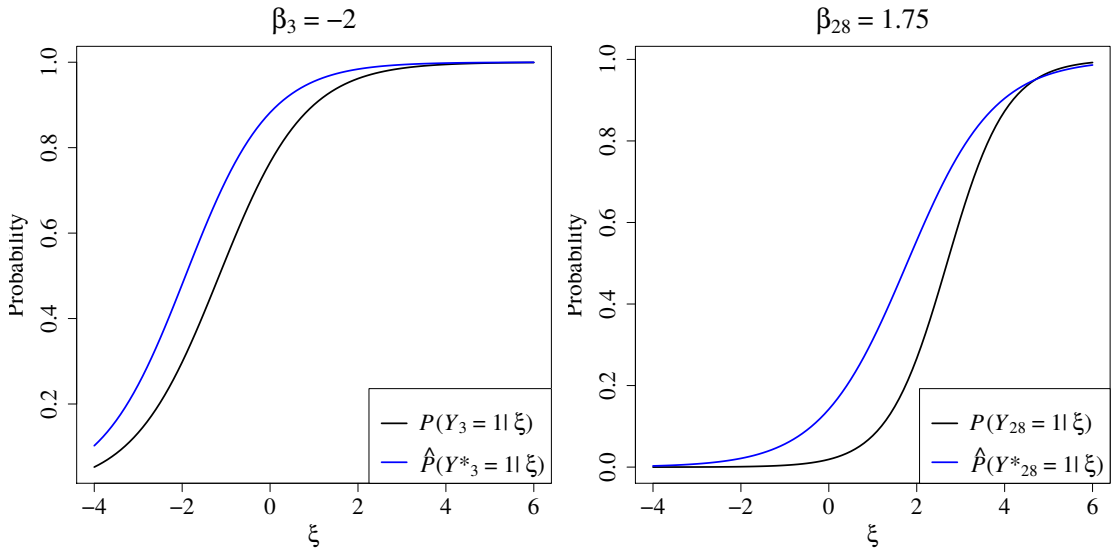

Figure 4.1: Graphical comparisons of $P(Y_i = 1|\xi)$ and $\hat{P}(Y_i^* = 1|\xi)$ for an exemplary item with low difficulty and low proportion of missing data (Y_3) and an exemplary item with high difficulty and high proportion of missing data (Y_{28}) using Data Example A.

estimates obtained of the 1PLM and the 2PLM using BILOG 3 with IAS. Columns two to four give the results under the treatment of missing data as wrong. Columns five to eight show the results under partial correct scoring (PCS) which will be discussed in the subsequent section. As expected, the item difficulties were overestimated for all items (see also Figure 4.2). The mean $\hat{\beta}_i = 1.145$ of the difficulty estimates is much higher

than the true mean $\bar{\beta} = -0.118$ ($t = 3.934$; $df = 29$, $p < 0.001$), erroneously indicating a considerably more difficult test. Using the 1PLM, the item fit measures indicated a bad model fit for all 30 items although Data Example A was generated using the Rasch model[5]. In real applications, the 2PLM could be chosen as a less restrictive alternative model in such a situation. The bias of item difficulty estimates is very close between 1- and 2PLM. As Figure 4.3 illustrates, the estimated item discriminations were increasingly overestimated the higher the proportion of missing responses per item was. The mean $\bar{\hat{\alpha}} = 1.206$ of the estimated discrimination parameters deviates significant from the true item discrimination $\alpha = 1$ ($t = 5.492, df = 29, p < 0.001$).

However, it is important to note that the estimates $\hat{\alpha}_i$ need not necessarily be positively biased when IAS is used for item nonresponses. In Data Example A it was assumed that the tendency to have item nonresponses is positively correlated with the latent ability ($Cor(\xi, \theta) = 0.8$). The bias might be different for other missing data mechanisms and other relations between the variables. Exemplarily, the case is examined were the missing data mechanism w.r.t. Y_i is MCAR. Thus the D_i is stochastically independent from Y_i and ξ, respectively. In this case it is still expected that the ICC would be right-shifted. Hence, the item difficulties are overestimated in presence of missing responses. However, the item discrimination is affected quite differently than for the case of non-ignorable missing data with $Cov(\xi, \theta) < 0$. This can be demonstrated studying the limits of $P(Y_i^* = 1 \,|\, \xi)$ given by

$$\begin{aligned} \lim_{\xi \to \infty} P(Y_i^* = 1 \,|\, \xi) &= \lim_{\xi \to \infty} P(Y_i = 1 \,|\, \xi) \cdot P(D_i = 1 \,|\, \xi) \\ &= \lim_{\xi \to \infty} P(Y_i = 1 \,|\, \xi) \cdot \lim_{\xi \to \infty} P(D_i = 1 \,|\, \xi). \end{aligned} \tag{4.16}$$

Equation 4.16 holds under any missing data mechanism considered in this work. If a latent variable ξ and a latent response propensity θ exist with $Cov(\xi, \theta) > 0$, and the regressions $P(Y_i = 1 \,|\, \xi)$ and $P(D_i = 1 \,|\, \theta)$ are monotonically increasing functions with the limits zero and one, then the upper limit under IAS is

$$\begin{aligned} \lim_{\xi \to \infty} P(Y_i^* = 1 \,|\, \xi) &= \lim_{\xi \to \infty} P(Y_i = 1 \,|\, \xi) \cdot \lim_{\xi \to \infty} P(D_i = 1 \,|\, \xi) \\ &= 1, \end{aligned} \tag{4.17}$$

[5]The χ^2-Test provided by BILOG indicated a significant deviation of the empirical ICCs from the model implied ICCs of the 1PLM for all items in Data Example A.

Table 4.1: Estimated item discriminations and item difficulties of the 1PLM and the 2PLM using IAS and PCS (Data Example A).

	IAS				PCS			
	1PLM		2PLM		1PLM		2PLM	
Item	α_i	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\beta}_i$	α_i	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\beta}_i$
Y_1	1	-1.597	0.974	-1.581	1	-1.669	0.863	-2.161
Y_2	1	-1.565	0.960	-1.565	1	-1.671	0.837	-2.215
Y_3	1	-1.170	1.029	-1.114	1	-1.405	0.804	-1.929
Y_4	1	-0.960	1.062	-0.895	1	-1.249	0.722	-1.871
Y_5	1	-0.334	1.034	-0.319	1	-1.007	0.603	-1.754
Y_6	1	-0.658	1.109	-0.596	1	-1.075	0.694	-1.664
Y_7	1	-0.321	1.130	-0.288	1	-0.917	0.622	-1.555
Y_8	1	-0.419	0.952	-0.424	1	-0.899	0.624	-1.520
Y_9	1	-0.306	0.990	-0.302	1	-0.766	0.657	-1.237
Y_{10}	1	-0.199	1.116	-0.181	1	-0.695	0.699	-1.066
Y_{11}	1	0.308	1.080	0.278	1	-0.566	0.495	-1.160
Y_{12}	1	0.265	1.037	0.244	1	-0.403	0.549	-0.748
Y_{13}	1	0.196	1.084	0.175	1	-0.444	0.599	-0.767
Y_{14}	1	0.682	1.260	0.564	1	-0.292	0.506	-0.575
Y_{15}	1	0.395	0.956	0.387	1	-0.172	0.562	-0.298
Y_{16}	1	0.776	1.052	0.718	1	-0.114	0.464	-0.223
Y_{17}	1	1.561	1.368	1.226	1	-0.124	0.276	-0.396
Y_{18}	1	1.243	1.194	1.060	1	-0.062	0.327	-0.146
Y_{19}	1	2.049	1.236	1.704	1	-0.021	0.170	-0.035
Y_{20}	1	2.237	1.342	1.768	1	0.048	0.146	0.427
Y_{21}	1	2.257	1.277	1.837	1	0.072	0.125	0.693
Y_{22}	1	2.645	1.507	1.947	1	0.105	0.096	1.235
Y_{23}	1	3.458	1.335	2.716	1	0.070	0.027	3.164
Y_{24}	1	2.914	1.422	2.212	1	0.111	0.090	1.389
Y_{25}	1	3.182	1.394	2.437	1	0.144	0.055	2.848
Y_{26}	1	2.807	1.249	2.310	1	0.244	0.040	6.327
Y_{27}	1	3.538	1.250	2.899	1	0.101	0.014	8.279
Y_{28}	1	3.687	1.765	2.455	1	0.189	1.490	2.627
Y_{29}	1	3.740	1.448	2.784	1	0.115	1.169	3.056
Y_{30}	1	3.950	1.561	2.803	1	0.136	1.109	3.170
Mean	1	1.145	1.206	0.842	1	-0.407	0.514	0.396

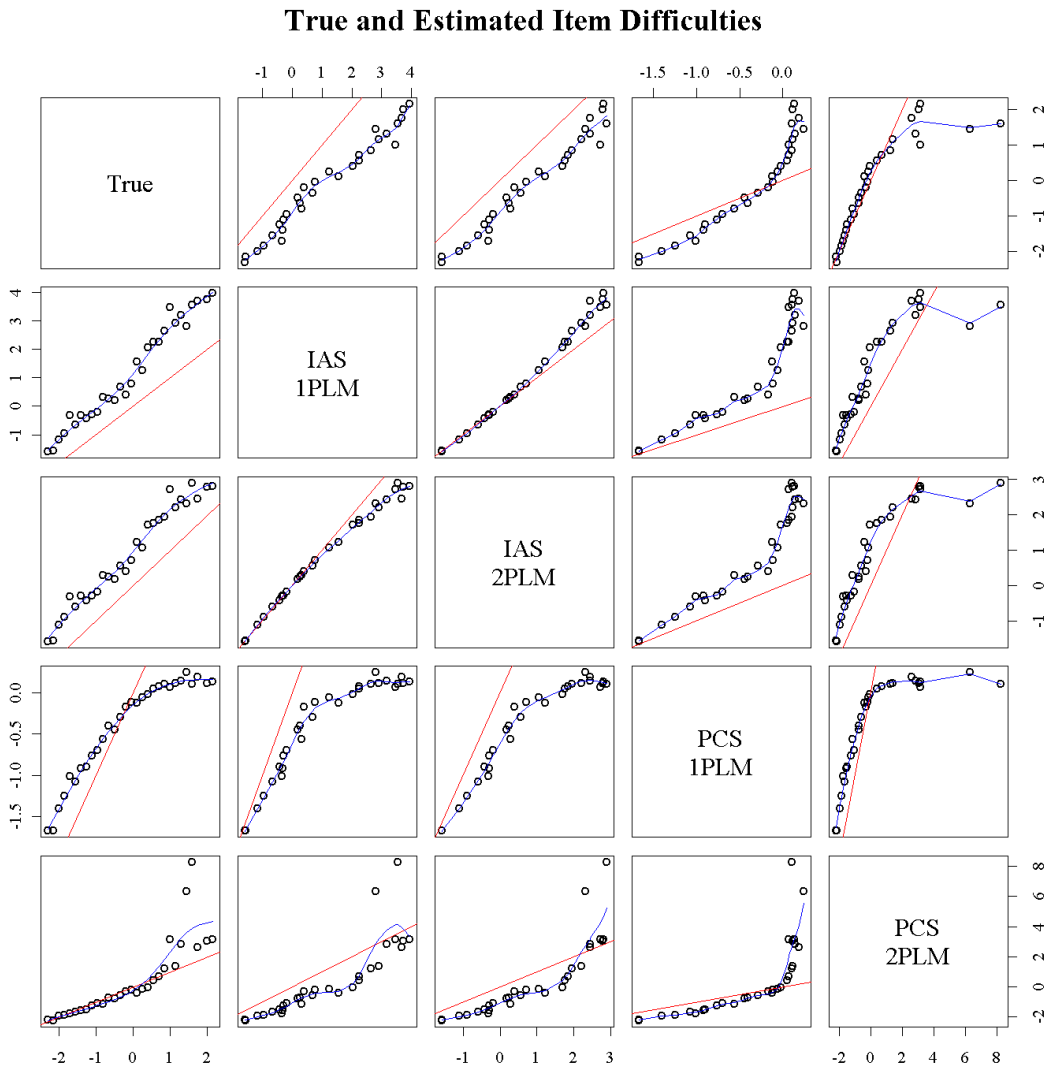

Figure 4.2: True and estimated item difficulties using IAS and PCS in the 1PLM and 2PLM. The red lines indicate the bisectric. The blue lines are smoothing spline regressions.

and the lower limit is

$$\begin{aligned} \lim_{\xi \to -\infty} P(Y_i^* = 1 \,|\, \xi) &= \lim_{\xi \to -\infty} P(Y_i = 1 \,|\, \xi) \cdot \lim_{\xi \to -\infty} P(D_i = 1 \,|\, \xi). \\ &= 0 \end{aligned} \tag{4.18}$$

Hence, using IAS the ICCs can also be described by a monotonically increasing function, with the lower asymptote equal to zero and the upper asymptote equal to one. However, if the missing data mechanism is missing completely at random, the upper limit of $P(Y_i^* =$

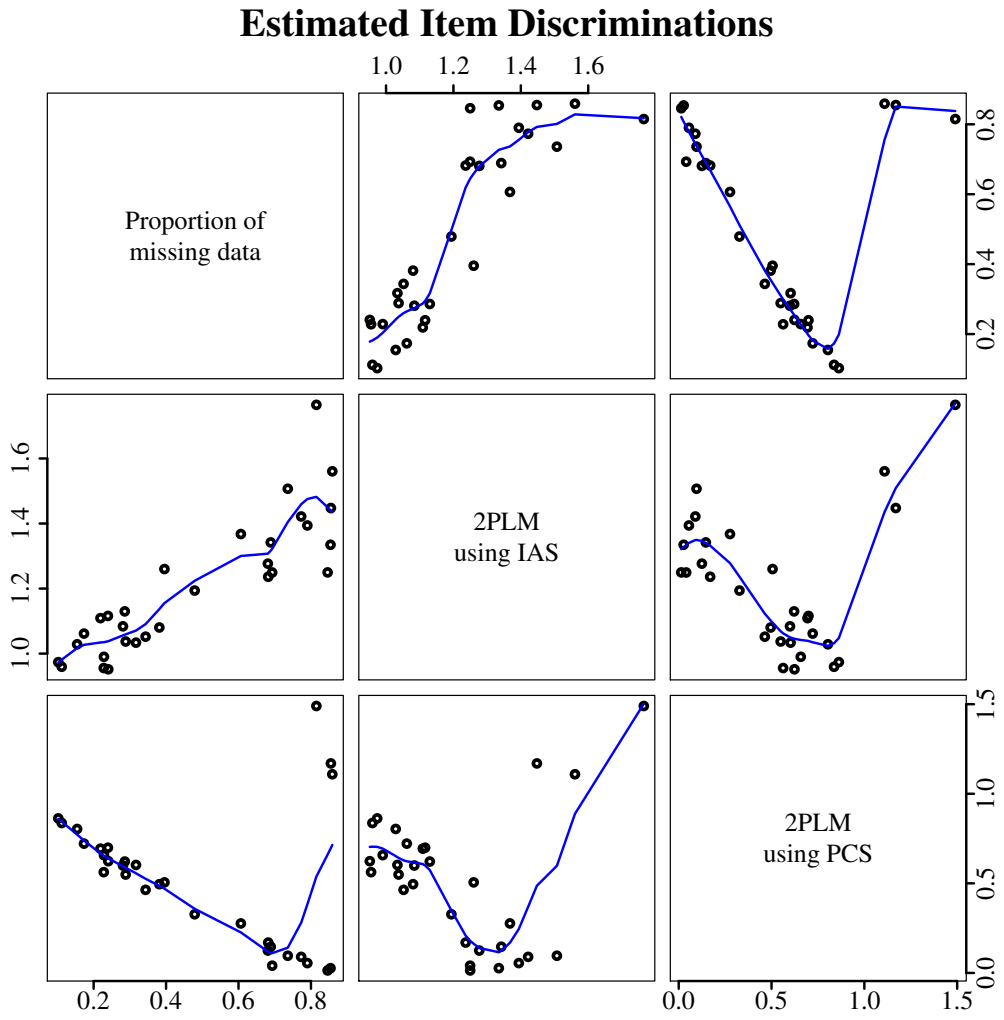

Figure 4.3: Relationship between item difficulties and estimated item discriminations when IAS and PCS is used in two-parameter models (Data Example A). The blue lines are smoothing spline regressions.

$1 \,|\, \xi)$ is

$$\begin{aligned} \lim_{\xi \to \infty} P(Y_i^* = 1 \,|\, \xi) &= \lim_{\xi \to \infty} P(Y_i = 1 \,|\, \xi) \cdot \lim_{\xi \to \infty} P(D_i = 1 \,|\, \xi) \\ &= 1 \cdot \lim_{\xi \to \infty} P(D_i = 1) \\ &= P(D_i = 1). \end{aligned} \tag{4.19}$$

Hence, the ICC of the variables Y_i^* cannot be described by 1-,2- or 3-parametric IRT models because the upper limits of these three IRT models is equal to one. Consequently, if

the missing data mechanism is MCAR and IAS is applied the measurement model will generally be miss specified using the 1-, 2-, or 3PLM. Neverthless, if these theoretical considerations are ignored and the 2- or 3PLM is used, then the item discrimination parameter will be negatively biased. To demonstrate this effect a single item with $\alpha = 1$ and $\beta = -2$ was simulated given the missing data mechanism is MCAR. The probability of observing any value of Y was $P(D = 1) = 0.7$. Figure 4.4 shows the results which are in line with the expectations derived theoretically. The red curve refers to a non-parametric binomial regression based on a local likelihood approach (Bowman & Azzalini, 1997). This curve approximates the true regression $P(Y_i^* = 1 \mid \xi)$ best. As Ramsey (1991) proposed, non-parametric ICC estimation is an appropriate model technique when parametric models fail to fit the data. The black curve of Figure 4.4 is obtained using the 2PLM. It can be seen that the item difficulty is overestimated ($\hat{\beta} = 0.692$) and the estimated item discrimination is considerably lower than one ($\hat{\alpha} = 0.348$). With increasing values of ξ, the non-parametric ICC approaches the theoretically implied upper limit of 0.7 (grey dotted line). The discrepancy between the non-parametric and the model-implied ICC of the 2PLM indicates that the Birnbaum-Model does not fit the filled-in data if IAS is used.

In summary, to treat missing data as incorrect responses leads not only to theoretical inconsistencies but also results in biased item parameter estimates. Whereas item difficulties are consistently overestimated depending on the proportion of missing data per item, the item discrimination parameter estimates might be biased either positively or negatively depending on the missing data mechanism and potentially many other factors. If the missing data mechanism is MCAR, then an upper asymptote ($P(D_i = 1)$) is implicitly introduced, which is incompatible with 1-, 2, and 3PL IRT models. However, if the 2- or 3PLM is erroneously applied, then the item discrimination parameters will be underestimated. Note that the aim of this investigation was not to study the bias of item parameters under all possible conditions but to demonstrate that IAS most likely results in biased parameter estimation in most real applications.

Effects of IAS on person parameter estimates Finally, the effect of IAS on person parameter estimates will be investigated. Table 4.2 shows the variances, covariances, and correlations between the true realized values of ξ underlying Data Example A and the ML estimates from BILOG 3 using IAS. The estimates of the 1- and 2PLM were compared. The differences between the MLEs of both models seem to be negligible although the relation seems to be nonlinear (see Figure 4.5). The correlation between the estimates is

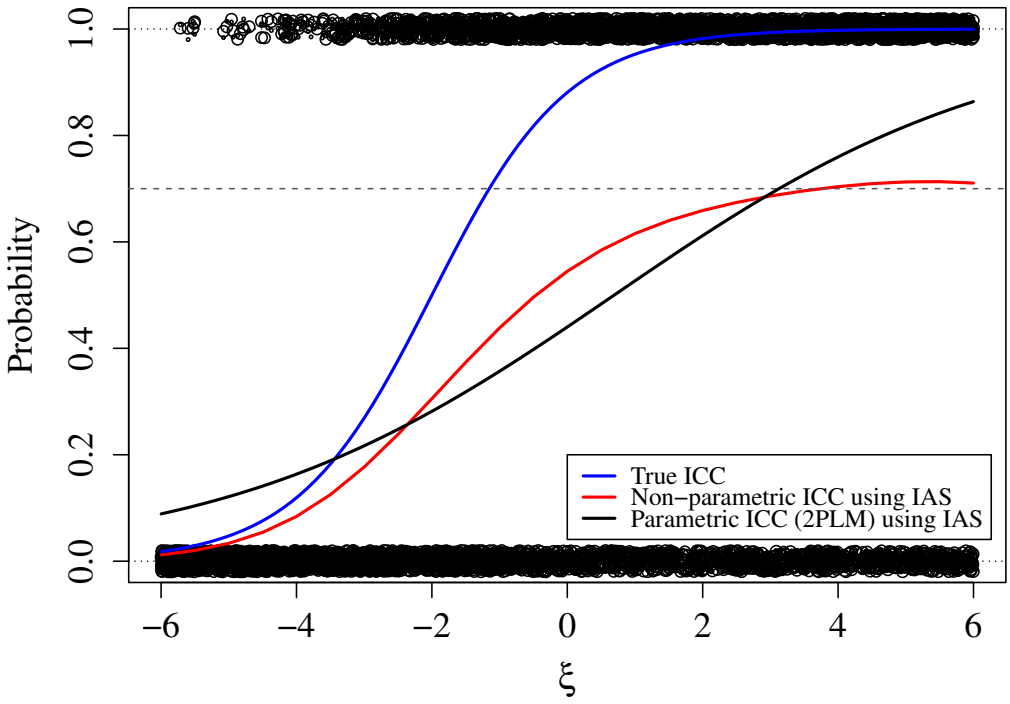

Figure 4.4: Effect of IAS on the estimation of parametric (2PLM) and non-parametric ICCs given the missing data mechanism is MCAR (true item parameters: $\alpha = 1$ and $\beta = -2$).

close to one. The MLEs of both models have approximately the same correlation ≈ 0.87 with the true values of ξ. This is slightly lower than the correlation between ξ and the ML estimates of the complete data ($r = 0.910$). This might imply that person parameter estimates are not affected. However, the model was identified by fixing the distribution of the latent variable to the values $E(\xi) = 0$ and $Var(\xi) = 1$ that were used for simulation. It is important to note that the item difficulties as locations of the latent variable considerably shifted. Hence, if item difficulties would be restricted to identify the model and the moments of the distribution of the latent variable would be freely estimated, then the results would be different. For example, if the mean of the item difficulties would be fixed to the true value $\sum_{i=1}^{I} \beta_i = -3.55$, then the distribution of the latent variable would be left-shifted. Hence, the person parameters would be underestimated. The point is that the item difficulties and the latent variable are shifted against each other. As a consequence, the item- and test information functions and, therefore, the standard errors differ. As discussed in Section 3.3, the functional form of item information functions $I_i(\xi)$, the test information function $\boldsymbol{I}(\xi)$, and the standard error function $SE(\xi)$ depend on item parameters α_i and β_i (see Equations 3.74 and 3.75). The overestimation of item difficulties

Table 4.2: Variances, Covariances and Correlations of the True Values of ξ and the MLE Estimates for Complete Data and the Filled-in Data Using IAS (Data Example A). Correlations are Marked by *.

	True	complete	IAS (1PLM)	IAS (2PLM)
True ξ (True)	1.002	0.910*	0.873*	0.868*
$\hat{\xi}_{ML}$ - complete data	1.041	1.307	0.886*	0.882*
$\hat{\xi}_{ML}$ - IAS (1PLM)	1.096	1.271	1.575	0.992*
$\hat{\xi}_{ML}$ - IAS (2PLM)	1.009	1.171	1.447	1.351

should result in a right-shifted test information function. For the case of positively biased item discrimination estimates, the test information function is potentially overestimated. Figure 4.6 shows the test information function and the standard error function based on item parameter estimates of the 2PLM when IAS and PCS are used. PCS will be discussed in the following section. As expected, the test information function is right-shifted. Due to overestimated item discriminations the test information is also overestimated in wide ranges of ξ. In application, one would mistakenly conclude that the test is more reliable in the upper range of ξ. In situations where the item calibration is used to establish item pools for computerized adaptive testing, this would be fatal. Especially if the missing data mechanism and/or the treatment of missing data are different between the item calibration and test application, then parameter estimation and standard errors can be biased. In the case of CAT, the item selection can be inefficient and the point estimation and standard errors can be biased. Biased item parameters and test information functions can also result in biased marginal reliability estimates. In Section 3.3 it was outlined that $Rel(\hat{\xi})$ can be interpreted as the average reliability over the distribution of the latent variable ξ. Thus, the value of the marginal reliability depends on the test information function and the distribution of ξ. Optimal values of $Rel^{(A)}(\hat{\xi})$ result if the probability density function of ξ and the test information function are proportional. Comparing the nonparametrically estimated densities of the ML estimates (see Figure 4.7) with the respective test information functions (see 4.6) reveals that the density of the latent variable and the test information function implied by item parameter estimates under IAS are not proportional. In contrast, the true density and the test information function based on true item parameters show that the test fit the distribution of the latent variable appropriately. This means that the test information function and the density function are approximately proportional. The location of the maximum of test information and the expected value $E(\xi)$ are almost equal. The item difficulties are optimally spread across the range of ξ. In this case the marginal relia-

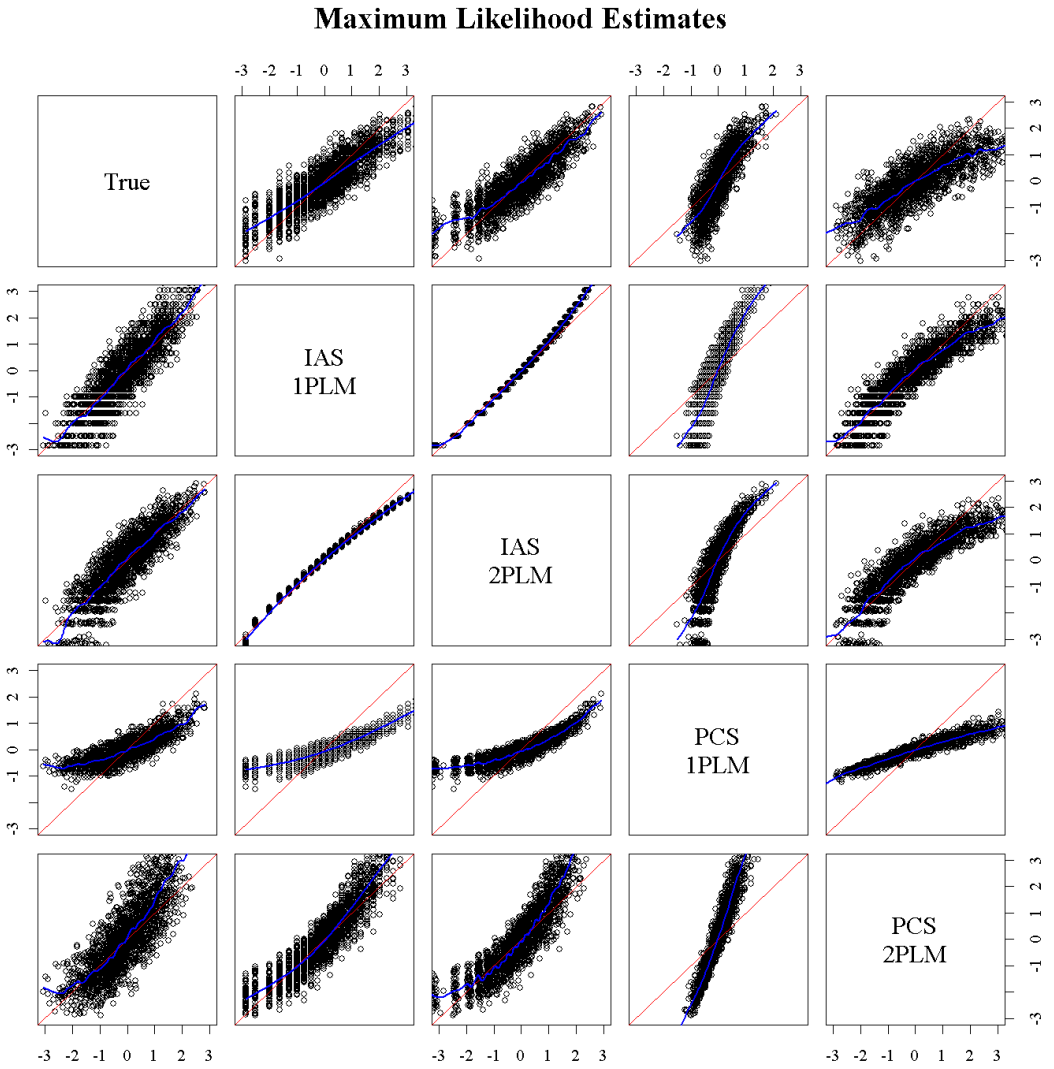

Figure 4.5: True person parameters compared to ML estimates in 1PL- and 2PL models when IAS and PCS are used. Red lines indicate the bisectric and blue lines represent smoothing spline regressions.

bility is close to the theoretical maximum. Using IAS, the maximum test information is in a range with a lower density and the marginal reliability is potentially underestimated with IAS. However, the results contradict the theoretical expectations. The marginal reliability in the 1PLM using the complete data was $Rel^{(A)}(\hat{\xi}_{ML}) = 0.835$. Using IAS, the marginal reliabilities were $Rel^{(A)}(\hat{\xi}_{ML}) = 0.824$ and $Rel^{(A)}(\hat{\xi}_{ML}) = 0.845$ in the 1PL- and 2PLM. The three coefficients are very close and the marginal reliability is nearly unaffected by IAS. This might be due to the overestimated test information function. However, since the

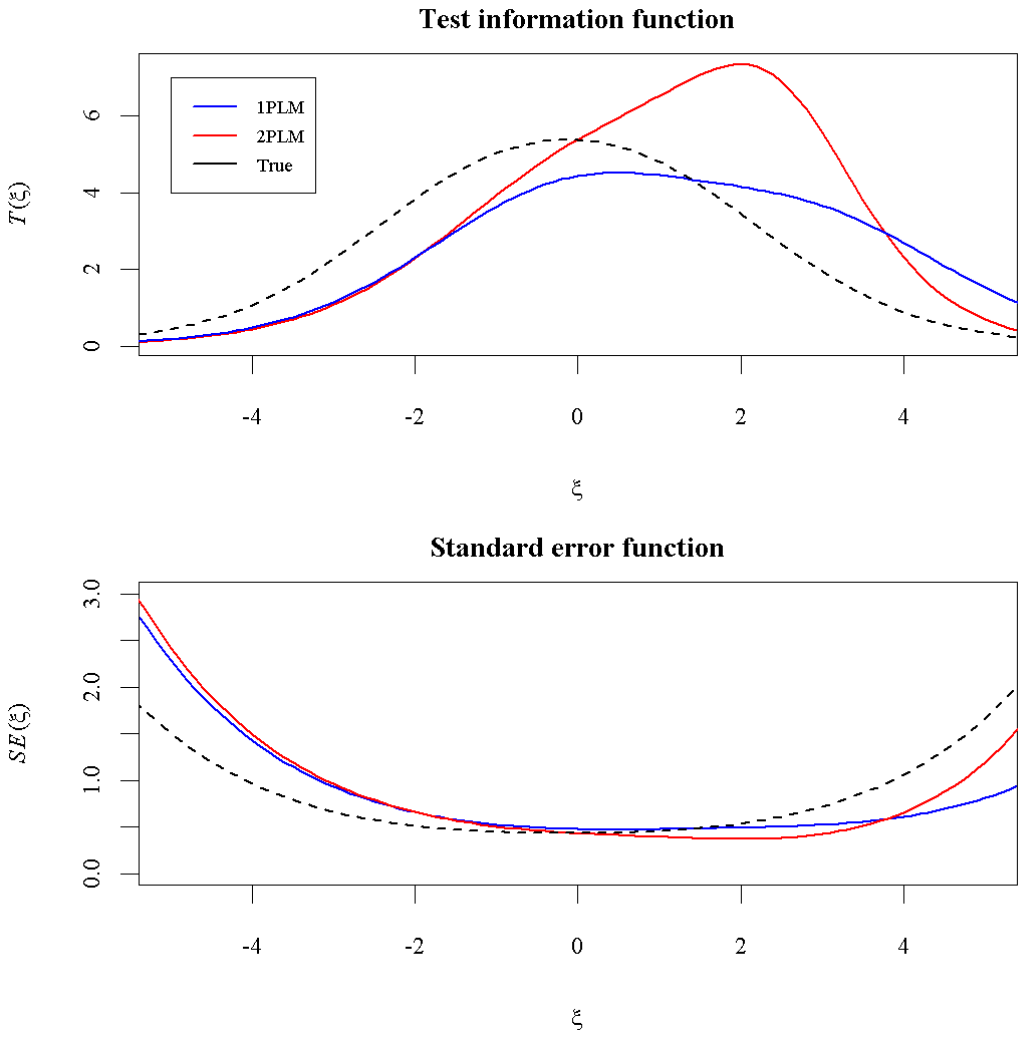

Figure 4.6: Estimated model-implied test information and standard error functions of the 1PLM and 2PLM using IAS.

latent variable seems to be estimated with a comparable accuracy for complete data and with missing data handled by IAS, the determination coefficients $R^2_{\hat{\xi}|\xi}$ of the regressions $E(\hat{\xi}\,|\,\xi)$ should be almost identical as well. In our simulated Data Example A we can use the true values of latent variable ξ and the estimates $\hat{\xi}$ of the different model to estimate the regression $E(\hat{\xi}\,|\,\xi)$ with $R^2_{\hat{\xi}|\xi} = Var[E(\hat{\xi}\,|\,\xi)]/Var(\hat{\xi})$ as an alternative estimate of the marginal reliability. Using the complete data $R^2_{\hat{\xi}|\xi}$ was 0.828, which is very close to the marginally reliability estimated in BILOG 3 ($Rel(\hat{\xi}_{ML}) = 0.835$). Using IAS, however, the determination coefficient was $R^2_{\hat{\xi}|\xi} = 0.762$ for the 1PLM and $R^2_{\hat{\xi}|\xi} = 0.753$ for the 2PLM. Both coefficients are lower than the marginal reliabilities of $\approx 0.82 - 0.85$. Such

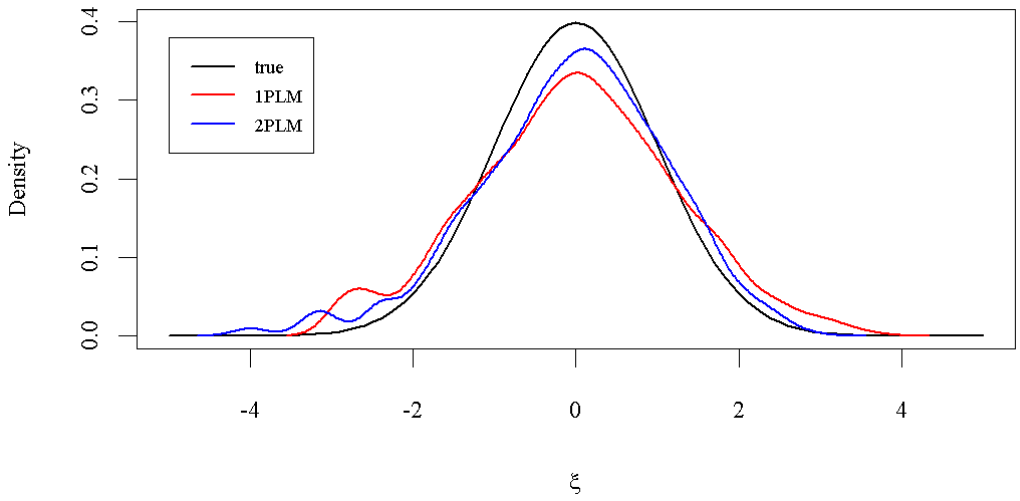

Figure 4.7: Non-parametrically estimated densities of ML person parameter estimates in the 1PLM and 2PLM using IAS.

a discrepancy between estimated marginal reliabilities $Rel(\hat{\xi})$ and the determination coefficients $R^2_{\hat{\xi}|\xi}$ under IAS has also been found for EAP person parameter estimates (Rose et al., 2010). It seems to be a consistent finding regardless of the type of the estimator (ML estimator or EAP). Here it is argued that the differences between $R^2_{\hat{\xi}|\xi}$ and the marginal reliabilities reflects the different construction of the latent variable ξ when IAS is used. As explained in detail at the beginning of this section, treating missing data as wrong means to replace the manifest variables Y_i in the measurement model with Y_i^*. The filled-in data set is treated as though no missing data would exist. Each value $y_i = 0$ can result from a wrong answer or a non-response to the item. As demonstrated above, the likelihood function does not distinguish between missing responses and incorrect answers. Therefore, the latent variables constructed in two measurement models using either the items $Y_1, \ldots, Y_I$ or $Y_1^*, \ldots, Y_I^*$ are potentially different. In order to distinguish between these two constructed latent variables, $\xi*$ denotes the latent variable in the measurement model based on Y_i^*. ξ remains the latent variable in the measurement model constituted by Y_i. As previously discussed for the sum score, the variable Y_i^* combines two pieces of information. Y_i^* is a function $f(Y_i, D_i)$ expressed by the assignment rule of Equation 3.7. Hence, information about performance with respect to the item Y_i and the willingness or ability to show this performance indicated by D_i are confounded into Y_i^*. From this point of view it can be expected that the latent variable ξ^* combines also information about a person?s ability and the tendency to respond to the items. To proof this hypothesis Data Example

A was used. The ML estimates obtained from the complete data (without missing data) and the ML estimates when IAS was applied to incomplete data were regressed on the true values of ξ and θ used in Data Example A. It was expected that the ML estimates $\hat{\xi}_{ML}$ based on the complete data are conditionally regressively independent from θ given ξ. In contrast, the estimates $\hat{\xi}^*_{ML}$ were expected to be conditionally regressively dependent on θ given ξ. The results are shown in Table 4.3 for the ML estimates obtained from the 1PLM and 2PLM. Six linear regression models were estimated: (a) the simple re-

Table 4.3: Regression Coefficients, $t-$ and $p-$values for Simple (SR) and Multiple Regressions (MR) of ML Person Parameter Estimates on the True Values of θ and ξ.

		Independent variables						
		ξ			θ			
MLE from	Model	Coeff.	t	p	Coeff.	t	p	R^2
Complete data	SR	0.910	97.98	< 0.001	/	/	/	0.828
Complete data	MR	0.924	60.02	< 0.001	0.018	-1.18	0.238	0.828
IAS (1PLM)	SR	0.873	79.89	< 0.001	/	/	/	0.762
IAS (1PLM)	MR	0.524	34.33	< 0.001	0.437	28.68	< 0.001	0.831
IAS (2PLM)	SR	0.868	77.98	< 0.001	/	/	/	0.753
IAS (2PLM)	MR	0.517	33.10	< 0.001	0.439	28.11	< 0.001	0.822

gressions (SR) $E(\hat{\xi}_{ML}\,|\,\theta)$ and $E(\hat{\xi}^*_{ML}\,|\,\theta)$; and (b) the multiple regressions $E(\hat{\xi}_{ML}\,|\,\xi,\theta)$ and $E(\hat{\xi}^*_{ML}\,|\,\xi,\theta)$. The results confirmed the theoretical expectations. The ML estimates $\hat{\xi}^*_{ML}$ are not conditionally regressively independent from θ given ξ. In contrast, the standardized partial regression coefficient of θ in the multiple regression (MR) of $\hat{\xi}_{ML}$ estimated from the complete data on (ξ,θ) is not significant. Additionally, the determination coefficients were higher in the multiple regressions $E(\hat{\xi}^*_{ML}\,|\,\xi,\theta)$ than in the simple regressions $E(\hat{\xi}^*_{ML}\,|\,\xi)$. However, θ does not explain additional variance in $\hat{\xi}_{ML}$ given the true values ξ. Note that in this data example the standardized partial regression coefficients of ξ and θ in the regressions $E(\hat{\xi}^*_{ML}\,|\,\xi,\theta)$ are pretty close. Hence, the latent variable ξ^* is a linear combination of the latent ability ξ and the latent response propensity θ. The meaning of the constructed latent variable using IAS is changed. ξ^* represents no longer the latent ability of interest but express both the latent ability and the tendency to respond to the test items. If one is only interested in the latent ability, construct irrelevant variance is introduced by IAS threatening the test fairness and the validity of the test scores. Thus, IAS is not appropriate to handle item nonresponses if the test is intended to estimate *only* the latent ability ξ. As the last column of Table 4.3 shows, the determination coefficients of the multiple regressions $E(\hat{\xi}^*_{ML}\,|\,\xi,\theta)$ are close to the marginal reliabilities $Rel(\hat{\xi}^*_{ML})$ esti-

mated by BILOG 3. Hence, the marginal reliabilities calculated based on item and person parameter estimates (without true values ξ) are not biased but are reliabilities of estimates $\hat{\xi}^*$ instead of $\hat{\xi}$.

In summary, to score item nonresponses as incorrect answers is associated with strong implicit assumptions about the missing data mechanism. IAS is incompatible with MCAR and MAR conditions. Even if the nonresponse mechanism is NMAR, IAS and the use of one- and two-parameter IRT models with standard ML estimation result in theoretical inconsistencies. Formally, this IAS means to replace the items $Y_1, \ldots, Y_I$ in the measurement model with $Y_1^*, \ldots, Y_I^*$. Implicitly, a different random experiment with a different probability space is considered. As a consequence, the item parameter estimates are biased in the sense that they do not estimate the parameters of the target model. The item difficulties are overestimated resulting in right-shifted ICCs. The item discrimination parameters in 2PLM might be affected differently, depending on the missing data mechanism. It was demonstrated that IAS introduces an upper asymptote if the missing data mechanism is MCAR. In this case none of the 1- to 3PL IRT models are suited to fit the data. If these models are nevertheless applied, then the item discrimination will be underestimated compared to the true item discrimination. In contrast, if the missing data mechanism is non-ignorable with a positive correlation between the latent ability and the latent response propensity, then the item discrimination will be overestimated. There might be many other conditions in presence of missing data not considered here that potentially result in other biased patterns. Finally, it could be shown that the latent variable is affected with respect to its meaning and interpretation. The variables $Y_1^*, \ldots, Y_I^*$ in the IAS measurement model contain two pieces of information: performance in the test and missingness. This is also reflected in the constructed variable ξ^* that is different from the latent ability ξ. The construction, the meaning, and, therefore, the interpretation of test scores using IAS differs. In this respect the treatment of missing data is also a matter of validity. The results show clearly that the implicit assumptions and implications of IAS are (a) not taken into account in the commonly used IRT models and in the ML estimation, and (b) not realistic for almost all real world applications. The treatment of missing data as wrong should be avoided.

4.3.2 Partially Correct Scoring of Item Nonresponses

IAS is a deterministic imputation model. As such it was often criticised (Lord, 1974, 1983a; Ludlow & O'Leary, 1999; Rose et al., 2010). Lord (1974, 1983a) stated that IAS is problematic using common ML estimation. This could be confirmed here. Lord

proposed two alternative approaches to handle missing responses in test items. Both account for the fact that test takers who omit an item Y_i have a non-zero probability to answer this item correctly even if a response is not observed. Lord suggested to make plausible assumptions about the mechanism that leads to missing data and to adapt the likelihood function accordingly. In his first approach introduced in 1974, Lord assumed that the probability $P(Y_i = 1 \,|\, U, D_i = 0)$ to answer item Y_i is a constant c. Typically c is chosen to be $1/A$ for multiple choice items with A as the number of response categories. Thus, it is assumed that each person would answer omitted items completely at random, so that each alternative is chosen with the same probability $c = P(Y_i = 1 \,|\, U, D_i = 0) = P(Y_i = 1 \,|\, D_i = 0)$. For dichotomous items that is $c = 0.5$. In this work PCS as proposed by Lord (1974) is also regarded as an imputation method for two reasons. First, in his original paper (1974) heproved that PCS is equivalent to the imputation of missing responses by random draws from a Bernoulli distributed random variable with $P(Y_i = 1) = c$ if $N \to \infty$. Second, the pseudo-likelihood $\mathcal{L}^\bullet(\boldsymbol{Y}; \boldsymbol{\iota})$ reveals that PCS means to impute c for missing values. The term pseudo-likelihood was used by Lord to underline the fact that the ML function used under PCS is not simply the likelihood of the observed data.

$$\begin{aligned} \mathcal{L}^\bullet(\boldsymbol{y}; \boldsymbol{\iota}) \quad = \quad & \prod_{n=1}^{N} \prod_{i=1}^{I} \Big\{ \Big[P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})^{y_{ni}} P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota})^{1-y_{ni}} \Big]^{d_{ni}} \\ & \cdot \Big[P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})^{c} P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota})^{1-c} \Big]^{1-d_{ni}} \Big\}. \end{aligned} \tag{4.20}$$

This estimation equation is nearly identical to the likelihood function $\mathcal{L}(\boldsymbol{y}^*; \boldsymbol{\iota})$ when IAS is used (see Equation 4.6). In fact, IAS can be seen as a special case of PCS with $c = 0$ for all items. Accordingly, some implications of PCS are similar to that of IAS, such as conditional stochastic independence $Y_i \perp U \,|\, D_i = 0$ and $Y_i \perp \xi \,|\, D_i = 0$. In fact, PCS assumes that the probability to solve a missing item is always c regardless of the test takers? ability and the difficulty of the items. Accordingly, the theoretical inconsistencies of the use of standard IRT models in conjunction with PCS and IAS are very similar. For example, the implications of conditional stochastic independence $Y_i \perp \xi \,|\, D_i = 0$ is ignored in Lord's pseudo-ML estimator. This can be shown considering the logarithm of the pseudo-likelihood $\ell^\bullet(\boldsymbol{Y}; \boldsymbol{\iota})$ under PCS that can be written as

$$\begin{aligned} \ell^\bullet(\boldsymbol{y}; \boldsymbol{\iota}) \quad = \quad & \sum_{n=1}^{N} \sum_{i=1}^{I} \Big\{ d_{ni} \cdot \left[y_{ni} \cdot P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota}) + (1 - y_{ni}) \cdot P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota}) \right] \\ & + (1 - d_{ni}) \cdot \left[c \cdot P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota}) + (1 - c) \cdot P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota}) \right] \Big\}. \end{aligned} \tag{4.21}$$

The first summand within the curly braces refers to the observed values $\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}}$. It can be seen that the responses $y_{ni} = 1$ or $y_{ni} = 0$ serve as selection variables. If item i was solved ($y_{ni} = 1$), then the conditional probability $P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})$ remains in the estimation equation. If item i was not solved ($y_{ni} = 0$), then the counter probability $P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota})$ is included. The second summand within the curly braces of Equation 4.21 refers to the missing responses $\boldsymbol{Y_{mis}} = \boldsymbol{y_{mis}}$ that are replaced by c. Since c is typically chosen to be greater than zero and, therefore, $1-c$ is lower than one, the constants c and $1-c$ act more as weights of $P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota})$ and $P(Y_{ni} = 0 \,|\, \xi; \boldsymbol{\iota})$ than selection variables. The consequences with respect to item and person parameter estimates need to be studied separately for each estimand. Here the analytical examination is confined to the person parameter estimation of a unidimensional latent variable ξ based on the response vector $\boldsymbol{Y_n} = \boldsymbol{y_n}$ using PCS. As introduced in Section 3.1.3, the ML estimate $\hat{\xi}_{ML}$ is found by maximizing the pattern log-likelihood $l(\boldsymbol{y_n}; \boldsymbol{\iota})$. Hence, $\hat{\xi}_{ML}$ is the value of the parameter space $\Omega_\xi = \mathbb{R}$ for which $\frac{\partial}{\partial \xi} \ell(\boldsymbol{y_n}; \boldsymbol{\iota}) = 0$. In application of PCS the first derivative of $\ell^\bullet(\boldsymbol{y_n}; \boldsymbol{\iota})$ is set to zero instead of $\ell(\boldsymbol{y_n}; \boldsymbol{\iota})$, with

$$\frac{\partial}{\partial \xi} \ell^\bullet(\boldsymbol{y_n}; \boldsymbol{\iota}) = \sum_{i=1}^{I} d_{ni} \cdot \alpha_i \cdot \left[y_{ni} - P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota}) \right] + (1 - d_{ni}) \cdot \alpha_i \cdot \left[c - P(Y_{ni} = 1 \,|\, \xi; \boldsymbol{\iota}) \right]. \tag{4.22}$$

This equation can be divided into two parts, so that

$$\frac{\partial}{\partial \xi} \ell^\bullet(\boldsymbol{y_n}; \boldsymbol{\iota}) = \underbrace{\sum_{i=1}^{I} d_{ni} \cdot \alpha_i \cdot \left[y_{ni} - P(Y_{ni} = 1 \,|\, \xi, \boldsymbol{\iota}) \right]}_{\frac{\partial}{\partial \xi} \ell^\bullet(\boldsymbol{y_{n;obs}}; \boldsymbol{\iota})} + \underbrace{\sum_{i=1}^{I} (1 - d_{ni}) \cdot \alpha_i \cdot \left[c - P(Y_{ni} = 1 \,|\, \xi, \boldsymbol{\iota}) \right]}_{\frac{\partial}{\partial \xi} \ell^\bullet(\boldsymbol{y_{n;mis}}; \boldsymbol{\iota})}. \tag{4.23}$$

The first part $\frac{\partial}{\partial \xi} \ell^\bullet(\boldsymbol{y_{n;obs}}; \boldsymbol{\iota})$ refers to the observed item responses and the second part $\frac{\partial}{\partial \xi} \ell^\bullet(\boldsymbol{y_{n;mis}}; \boldsymbol{\iota})$ refers to the missing responses. Under common regularity conditions certain properties of ML estimates follow without further assumptions. For example, the expectation $E(\frac{\partial}{\partial \xi} \ell(\boldsymbol{Y}; \boldsymbol{\iota})) = 0$ is always zero (Green, 2012). Otherwise, the ML estimator would be biased. In the case of person parameter estimation that means that the expected value $E(\frac{\partial}{\partial \xi} \ell(\boldsymbol{Y_n}; \boldsymbol{\iota}))$ needs to be zero for each test taker n in a sample of $n = 1, \ldots, N$. For brevity, only the single unit trial is considered in the further derivations. Hence, the subscript n can be omitted. Given the measurement model is correctly specified it follows that the conditional expectation $E(\frac{\partial}{\partial \xi} \ell(\boldsymbol{Y}; \boldsymbol{\iota}) \,|\, \xi) = 0$. This means that if a test could be

repeated infinitely to a single person with a particular fixed ability level, then the mean of the first derivatives of the response pattern likelihoods would be zero. Mathematically, that is,

$$\begin{aligned} E\left(\frac{\partial}{\partial\xi}\ell(\boldsymbol{Y};\boldsymbol{\iota})\,\middle|\,\xi\right) &= E\left(\sum_{i=1}^{I}\alpha_i\cdot[Y_i - P(Y_i=1\,|\,\xi;\boldsymbol{\iota})]\,\middle|\right) \\ &= \sum_{i=1}^{I}\alpha_i\cdot E\left[Y_i - P(Y_i=1\,|\,\xi;\boldsymbol{\iota})\,|\,\xi\right] \\ &= \sum_{i=1}^{I}\alpha_i\cdot[E(Y_i\,|\,\xi) - P(Y_i=1\,|\,\xi;\boldsymbol{\iota})] = 0, \end{aligned} \tag{4.24}$$

since $E(Y_i\,|\,\xi) = P(Y_i = 1\,|\,\xi;\boldsymbol{\iota})$ given the measurement model is correctly specified[6]. This property applies to each nonempty subset of items implying that $E(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{obs}};\boldsymbol{\iota})\,|\,\xi,\boldsymbol{D})$ should be zero as well. Note that $\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}}$ is the response vector of that subset of items which test taker n has answered, indicated by the missing indicator vector $\boldsymbol{D} = \boldsymbol{d}$. In other words, the expectation of the first derivative should be zero given the ability and *for each* missing pattern $\boldsymbol{D} \neq \boldsymbol{0}$. The conditional expectation $E(\frac{\partial}{\partial\xi}\ell(\boldsymbol{Y};\boldsymbol{\iota})\,|\,\xi,\boldsymbol{D})$ is equal to the conditional expectation of Equation 4.23 given $(\xi,\boldsymbol{\theta})$. That is,

$$\begin{aligned} E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) &= E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{obs}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) + E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{mis}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) \\ &= E\left(\sum_{i=1}^{I}D_i\cdot\alpha_i\cdot[Y_i - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})]\,\middle|\,\xi,\boldsymbol{D}\right) \\ &\quad + E\left(\sum_{i=1}^{I}(1-D_i)\cdot\alpha_i\cdot[c - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})]\,\middle|\,\xi,\boldsymbol{D}\right). \end{aligned} \tag{4.25}$$

The first conditional expectation can be written as

$$E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{obs}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) = \sum_{i=1}^{I}E\left(D_i\cdot\alpha_i\cdot[Y_i - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})]\,|\,\xi;\boldsymbol{D}\right) \tag{4.26}$$

[6]The functional form of the regression $P(Y_i = 1\,|\,\xi;\boldsymbol{\iota})$ determined by the item parameters in $\boldsymbol{\iota}$ needs to be correct. In this case the parametric regression $P(Y_i = 1\,|\,\xi;\boldsymbol{\iota})$ is equal to the true regression $E(Y_i\,|\,\xi)$.

Since $D_i = f(\boldsymbol{D})$, it follows that

$$\begin{aligned} E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{obs}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) &= \sum_{i=1}^{I}\alpha_i\cdot D_i\cdot E[Y_i - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})\,|\,\xi,\boldsymbol{D}] \\ &= \sum_{i=1}^{I}\alpha_i\cdot D_i\cdot\left[E(Y_i\,|\,\xi,\boldsymbol{D}) - P(Y_i=1\,|\,\xi;\boldsymbol{\iota})\right]. \end{aligned}$$

If no DIF exists so that $Y_i \perp \boldsymbol{D}\,|\,\xi$, then

$$E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{obs}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) = \sum_{i=1}^{I}\alpha_i\cdot D_i\cdot\left[E(Y_i\,|\,\xi) - P(Y_i=1\,|\,\xi;\boldsymbol{\iota})\right] = 0, \quad (4.27)$$

given $E(Y_i\,|\,\xi) = P(Y_i = 1\,|\,\xi;\boldsymbol{\iota})$. The latter holds if the measurement model is correctly specified. From these derivations it follows that the mean of the first derivatives of the pattern likelihoods of a person with a given ability level and for each missing pattern $\boldsymbol{D} = \boldsymbol{d}$ is zero. Hence, the person parameters would be estimated unbiasedly using the observed item responses if the true item parameters ($\boldsymbol{\iota}$) are known. However, the conditional expectation of $\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{mis}};\boldsymbol{\iota})$ given $(\xi,\boldsymbol{D})$ includes the constants c. For all $\boldsymbol{D} \neq \boldsymbol{1}$ it follows that

$$E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{mis}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) = \sum_{i=1}^{I}E\left[(1-D_i)\cdot\alpha_i\cdot\left[c - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})\right]\,|\,\xi;\boldsymbol{D}\right] \quad (4.28)$$

Since $(1 - D_i) = f(\boldsymbol{D})$ that is,

$$\begin{aligned} E\left(\frac{\partial}{\partial\xi}\ell^{\bullet}(\boldsymbol{Y_{mis}};\boldsymbol{\iota})\,\middle|\,\xi,\boldsymbol{D}\right) &= \sum_{i=1}^{I}(1-D_i)\cdot\alpha_i\cdot E\left[c - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})\,|\,\xi,\boldsymbol{D}\right] \quad (4.29) \\ &= \sum_{i=1}^{I}(1-D_i)\cdot\alpha_i\cdot[c - P(Y_i=1\,|\,\xi,\boldsymbol{\iota})]. \end{aligned}$$

This expression is not necessarily equal to zero if at least one item has been answered, implying that the person parameter estimates using PCS are potentially biased. The reason is that the weighted sum of the differences $c - P(Y_i = 1\,|\,\xi,\boldsymbol{\iota})$ appears in the likelihood function, which is inconsistent with the assumption that test takers who omit items are completely undecided about the correct answer. In other words, PCS assumes that test takers would answer omitted items independently of their ability ξ by pure guessing. Interestingly, this assumption implies that responses to missing items are absolutely non-

informative with respect to the latent ability ξ. Furthermore, this assumption requires that the differences $c - P(Y_i = 1 \,|\, \xi, \boldsymbol{\iota})$ needs to be replaced by $c - 1/A$ to yield the correct estimation function. However, since c is chosen to be equal to $1/A$ this difference is always zero. Hence, the part $\frac{\partial}{\partial \xi}\ell^{\bullet}(\boldsymbol{Y_{mis}}; \boldsymbol{\iota})$ of the estimation equation would be a constant that does not contribute to any estimand of the target model. In summary, the estimation function used in conjunction with PCS is inconsistent with the underlying assumption $Y_i \perp \xi \,|\, D_i = 0$. This assumption implies also that responses to missing items would not contribute to parameter estimation.

So far, it was demonstrated that essential properties of the log-likelihood function and their derivatives change if PCS is used. This implies biased ML parameter estimation. Although only scrutinized for person parameter estimation, ML estimation of item parameters can be shown to be biased as well. The differences $Y_i - P(Y_i = 1 \,|\, \xi, \boldsymbol{\iota})$ and $c - P(Y_i = 1 \,|\, \xi, \boldsymbol{\iota})$ appear also in the estimation equations of the item parameters α_i and β_I. Hence, the expected values $E(\frac{\partial}{\partial \alpha_i}\ell^{\bullet}(\boldsymbol{Y_{mis}}; \boldsymbol{\iota}))$ and $E(\frac{\partial}{\partial \beta_i}\ell^{\bullet}(\boldsymbol{Y_{mis}}; \boldsymbol{\iota}))$ evaluated at the true values of item and person parameters are also different from zero, implying that item parameter estimates are generally biased if PCS is used to handle item nonresponses. In the next step it will be further examined how item and person parameters are biased starting with an extreme example of a person u who is totally unwilling to answer any item. Hence, $\boldsymbol{d} = \mathbf{0}$. The first derivative of the pseudo-likelihood function of the completely unobserved response vector is $\frac{\partial}{\partial \xi}\ell^{*}(\boldsymbol{Y}; \boldsymbol{\iota}) = \frac{\partial}{\partial \xi}\ell^{\bullet}(\boldsymbol{y_{mis}}; \boldsymbol{\iota})$, which is given by

$$\begin{aligned} \frac{\partial}{\partial \xi}\ell^{\bullet}(\boldsymbol{y_{mis}}; \boldsymbol{\iota}) &= \sum_{i=1}^{I}(1 - d_i) \cdot \alpha_i \cdot [c - P(Y_i = 1 \,|\, \xi; \boldsymbol{\iota})] \\ &= \sum_{i=1}^{I} \alpha_i \cdot c - \sum_{i=1}^{I} \alpha_i \cdot P(Y_i = 1 \,|\, \xi; \boldsymbol{\iota}). \end{aligned} \tag{4.30}$$

This difference is set equal to zero in order to estimate the person's latent ability. If the items Y_i are dichotomous, then $c = 0.5$. In the case of the Rasch model $\alpha_i = 1$, for all $i = 1, \ldots, I$. The minuend of Equation 4.30 is then $0.5 \cdot I$. In other words, the person without any item response is assumed to have 50% correct item responses, regardless of the difficulties of the test items and the proficiency levels of the test taker. The latent ability is estimated, so that the weighted sum of the regressions - the subtrahend of Equation 4.30 - is also $0.5 \cdot I$. Hence, for persons with low ability levels it is potentially beneficial to omit difficult items, whereas highly proficient persons are expected to be penalized by PCS, especially if the omitted items are easy. The fundamental problem is that the constant

c is treated in the same way as an observed item response y_i that results from cognitive processing based on ξ. This contradicts the key assumption explicitly made by Lord that examinees would respond completely at random to the omitted items if they were required to answer. Standard IRT models do not account for this assumption and are thus inappropriate. The expected biases of person parameter estimates will be illustrated by Data Example A. Recall that it is expected that especially persons with lower ability levels will profit from PCS. In Data Example A the correlation $Cor(\xi, \theta)$ between the latent ability and the latent response propensity was 0.8. Therefore, the proportions of missing responses decreased with higher proficiency levels. The lower the proportion of item nonresponses are, the lesser PCS should affect person parameter estimation. Hence, the expected negative bias in higher ability levels should be small in Data Example A. Figure 4.8 (left) shows the person parameter estimate obtained from estimating ξ based on true item parameters. As expected, especially low proficient persons profit from omissions of items. If the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MCAR (implied by $Cor(\xi, \theta) = 0$),

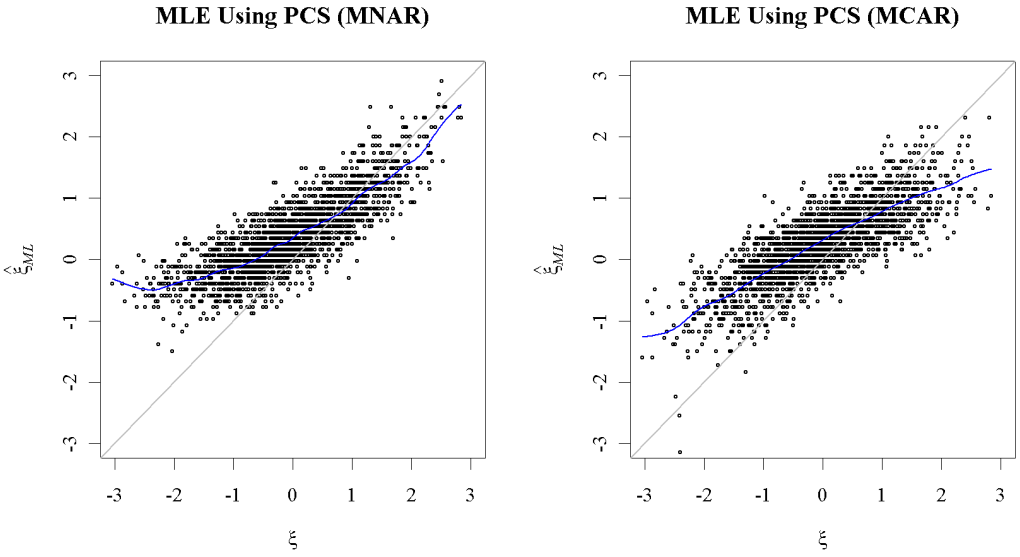

Figure 4.8: Comparison of true person parameters and ML person parameter estimates when PCS was used. Results are displayed for nonignorable missing data (left) and missing data that are MCAR (right). The grey lines are the bisectric. The blue lines are smoothing spline regressions.

then the probability of item nonresponses is the same for all ability levels. In this case the expected negative bias in person with high value of ξ could be confirmed as well. In Figure 4.8 (right) the person parameter estimates are compared with the true values of ξ

when the missing data mechanism is MCAR. This data example was simulated using the same parameters as in Data Example A except for $Cor(\xi, \theta)$ which was chosen 0.8[7]. The bias of is on average positive for lower ability levels and negative for higher values of ξ. A considerable shrinkage of the ML estimates results. The variance of the estimates in Data Example A was merely 0.391 and $s^2(\hat{\xi}) = 0.403$ for the Data Example A in the right graph of Figure 4.8. So far, the true item parameters were assumed to be known. Typically, they need to be estimated form the data as well. The effect of PCS on item parameter estimation will be examined next.

Impact of PCS on Itemparameter estimation In his original paper Lord proofed mathematically that PCS is equivalent to the imputation of random draws from a Bernoulli distributed random variable with $P(Y_i = 1 \mid D_i = 0) = c$ if $N \to \infty$. Interestingly, this proof implies systematic bias of item parameter estimates if PCS is used for item nonresponses. The random draws are stochastically independent of the test taker's ability. Strictly speaking, noise is imputed into the observed data. Accordingly, the sample estimates of the correlation between item responses to item i and ξ should decrease with higher proportions of missing data on item i. The item discrimination parameters α_i quantify the strength of the stochastic dependencies between the items Y_i and the latent variable. Hence, the sample estimates $\hat{\alpha}_i$ are expected to be systematically underestimated. The negative bias should increase, the higher the proportion of item nonresponses is. If all responses to item i are missing, the item vector consists of N repetitions of the constant c. In this case $\hat{\alpha}_i = 0$.

Table 4.1 presents the item parameter estimates of Data Example A obtained with PCS. The item difficulties are differently biased depending on whether the 1PL- or the 2PLM was applied. The item difficulty estimates have a non-linear relation to the true parameters β_i using the 1PLM (see Figure 4.2). Easier items have positively biased difficulty estimates, whereas difficult items show negatively biased estimates $\hat{\beta}_i$. Using the 2PLM, the estimates $\hat{\beta}_i$ of the easier items Y_1 to Y_{22} are pretty close to the true parameters. The more difficult items with higher proportions of missing data are severely overestimated. Some of these items (Y_{28} - Y_{30}) show also extremely distorted item discrimination estimates (see Figure 4.3), which may indicate numerical problems in the estimation procedure. Apart from these items, the bias of the item discrimination estimates shows exactly the expected bias: The underestimation of α_i increases with higher proportions of missing responses. For large proportions of missing data, $\hat{\alpha}_i$ tend toward zero caused by the im-

[7]The item parameters used for the the data example in the right graph of Figure 4.8 are given in Table 3.1. The overall proportion of missing responses was 48%.

putation of the constant $c = 0.5$. The mean of the estimated item discrimination is merely $\bar{\hat{\alpha}}_i = 0.397$, which is significantly different from the true value $\bar{\alpha}_i = 1$, ($t = -7.1653$, $df = 29$, $p < 0.001$).

In real applications using MML estimation, the item parameter estimates are typically used for subsequent person parameter estimation. Biased item parameter estimates most likely result in biased person parameter estimates. Figure 4.5 shows the ability estimates obtained from PCS using the 1PL- and 2PLM in comparison to IAS and the true values of ξ. In fact, a curvilinear stochastic relation could be found between the true values of ξ and their estimates $\hat{\xi}_{PCS}$. A R^2-difference test indicated that the regression model $E(\hat{\xi}_{PCS} \mid \xi) = \beta_0 + \beta_1\xi + \beta_2\xi^2$ fits the data significantly better than a linear regression $E(\hat{\xi}_{PCS} \mid \xi) = \alpha_0 + \alpha_1\xi$ (1PLM: $R^2_{dif.} = 0.028$, $F = 186.13$, $df = 1$, $p < 0.001$; 2PLM: $R^2_{dif.} = 0.018$, $F = 124.28$, $df = 1$, $p < 0.001$). If the Rasch model is applied in combination with PCS, the item discriminations are forced to be equal to one resulting in biased difficulty estimates $\hat{\beta}_i$ and person parameter estimates $\hat{\xi}_{PCS}$. Especially the variance was remarkably reduced, ($s^2(\hat{\xi}_{PCS}) = 0.244$). In comparison, the variance was $s^2(\hat{\xi}_{PCS}) = 2.258$ when the 2PLM was applied in conjunction with PCS. However, the person parameter estimates from both models - 1- and 2PLM - are highly correlated ($r = 0.954$). Table summarizes the variances, covariances, and correlations between the estimates $\hat{\xi}_{PCS}$ and the true values ξ underlying Data Example A.

Table 4.4: Variances, Covariances and Correlations of True Values ξ and ML Estimates of Complete Data and Filled-in Data Using PCS (Data Example A). Correlations are marked by *.

	True	$\hat{\xi}_{ML}$ - complete	$\hat{\xi}_{PCS}$ - 1PLM	$\hat{\xi}_{PCS}$ - 2PLM
ξ - true	1.002	0.910*	0.823*	0.825*
$\hat{\xi}_{ML}$ - complete data	1.041	1.307	0.897*	0.875*
$\hat{\xi}_{ML}$ - PCS (1PLM)	0.407	0.506	0.244	0.954*
$\hat{\xi}_{ML}$ - PCS (2PLM)	1.240	1.503	0.708	2.258

Biased item parameter estimates affects also the functional form of the test information function $\boldsymbol{I}(\xi)$ and the standard error function $SE(\xi)$, respectively. Figure 4.9 shows the different functions $\boldsymbol{I}(\xi)$ implied by the item parameter estimates of 1- and 2PLM obtained from Data Example A using PCS. The peaked test information function in the Rasch model resulted from the strong shrinkage of the item difficulty estimates that ranged merely between -1.669 and 0.244. Recall that the true item difficulties were chosen between -2.30 - 2.15. In turn, the low test information function in 2PLM is caused by the strongly negatively biased estimates $\hat{\alpha}_i$. The marginal reliabilities estimated in

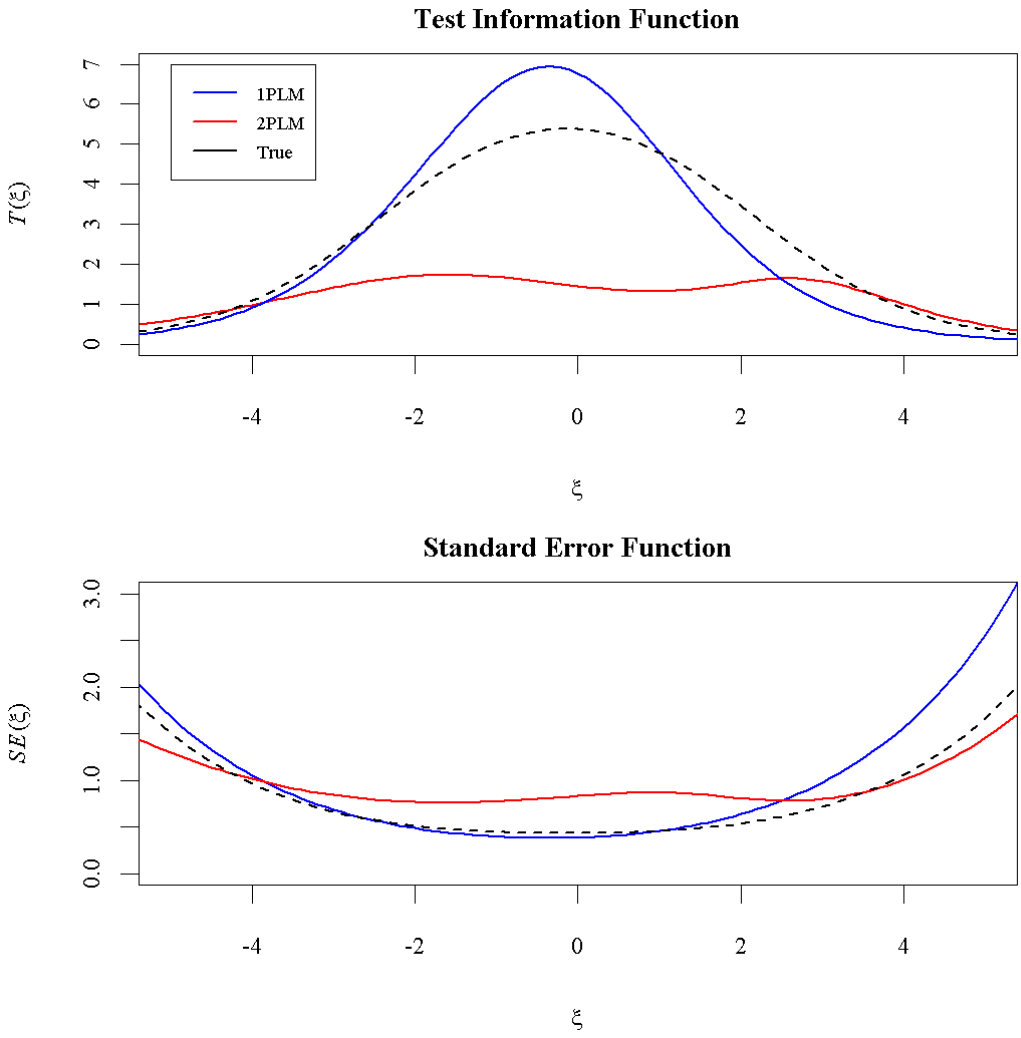

Figure 4.9: Estimated model-implied test information and standard error function of the 1PLM and 2PLM using PCS.

BILOG 3 were both very low: $Rel(\hat{\xi}_{PCS}) = 0.368$ (1PLM), and $Rel(\hat{\xi}_{ML}) = 0.5681$ (2PLM). This is even far below the squared correlations $r^2_{1PLM}(\xi, \hat{\xi}_{PCS}) = 0.677$ and $r_{1PLM}(\xi, \hat{\xi}_{PCS}) = 0.680$. In other words, the marginal reliabilities estimated in conjunction with PCS are also not trustworthy and should not be interpreted. Of course, due to the systematic bias implied by the non-linear relation between the estimates $\hat{\xi}_{PCS}$ and ξ this fact is of minor importance.

Finally, the impact of PCS with respect to the construction of the latent variable was examined by means of regression analyses. The estimates $\hat{\xi}_{PCS}$ from Data Example A were regressed on the latent variables ξ and θ. If the estimator is unbiased, then the

regression should be linear with an intercept equal to zero and the regression coefficient of ξ equal to one. Additionally, $\hat{\xi}$ should be stochastically independent of θ given ξ, implying that the regression coefficient of θ in a multiple regression $E(\hat{\xi}_{PCS} \mid \xi, \theta)$ should be zero. This was found for the ML person parameter estimates of the complete data (see Table 4.2). For the case of IAS it could be demonstrated that the person parameter estimates are not regressively independent of θ given ξ. The results imply that the latent variable constructed using IAS is a linear combination of ξ and θ and not simply ξ. The effects on item parameter estimates were quite different between IAS and PCS the same might be true with respect to the construction of the latent variable. In contrast to IAS, missing response are not scored as wrong answers. Considering the filled-in data set under PCS it can be distinguished whether the values result from completed items ($y_i = 0$ or $y_i = 1$) or from item nonresponses $y_i = c$. However, in the estimation procedures the imputed values $y = c$ are treated as regular responses as though test takers had proceeded the item due to their ability. Similar to IAS, two pieces of information are mixed-up in the filled-in data set using PCS: (a) performance in the test, and (b) willingness or ability to provide a response. It is expected that the latent variable ξ_{PCS} constructed using PCS might reflect this counfounding.

A multiple linear regression model was chosen to estimate the parameters of $E(\hat{\xi}_{PCS} \mid \xi, \theta)$. The non-linear relationship found between $\hat{\xi}_{PCS}$ and ξ was taken into account by including the squared variables ξ^2 and θ^2. Additionally, the interaction term $\xi \cdot \theta$ was included. An interaction between ξ and θ with respect to $\hat{\xi}_{PCS}$ is very likely since a quadratic relationship between $\hat{\xi}_{PCS}$ and ξ is implied if (a) $Cov(\xi, \theta) \neq 0$, and (b) an interaction between ξ and θ exists [8]. The results of the regression analyses are given in Table 4.5. Two regression models were applied with (a) the estimates $\hat{\xi}_{PCS}$ obtained from the 1PLM and (b) from the 2PLM. In both regressions the person parameter estimates were found to be stochastically dependent on θ given ξ. As expected, there was a significant interaction effect between the latent variable ξ and θ with respect to $\hat{\xi}_{PCS}$. The contribution of the quadratic term ξ^2 is relatively small. The regression coefficient of the conditional regression of the estimator $\hat{\xi}_{PCS}$ on its estimand ξ is moderated by the latent response propensity θ. Recall that the person parameter estimates were increasingly positively biased by PCS the lower the latent ability is (see Figures 4.5 and 4.8).

A multiple linear regression model was chosen to estimate the parameters of $E(\hat{\xi}_{PCS} \mid \xi, \theta)$. The non-linear relationship found between $\hat{\xi}_{PCS}$ and ξ was taken into account by including the squared variables ξ^2 and θ^2. Additionally, the interaction term $\xi \cdot \theta$ was included. An interaction between ξ and θ with respect to $\hat{\xi}_{PCS}$ is very likely since a quadratic relation-

[8] Since $E(Y \mid X, Z) = E[E(Y \mid X, Z) \mid X]$. In a linear multiple regression that is, $E(\alpha_0 + \alpha_1 X + \alpha_2 Z + \alpha_3 XZ \mid X) = \alpha_0 + \alpha_1 X + E(\alpha_2 Z \mid X) + \alpha_3 E(Z \mid X)X$. Let $E(Z \mid X) = \beta_0 + \beta_2 X$, then $E(Y \mid X, Z) = (\alpha_0 + \alpha_2\beta_0) + (\alpha_1 + \alpha_2\beta_1 + \alpha_3\beta_0)X + \alpha_3\beta_1 X^2$.

ship between $\hat{\xi}_{PCS}$ and ξ is implied if (a) $Cov(\xi, \theta) \neq 0$, and (b) an interaction between ξ and θ exists [9]. The results of the regression analyses are given in Table 4.5. Two regression models were applied with (a) the estimates $\hat{\xi}_{PCS}$ obtained from the 1PLM and (b) from the 2PLM. In both regressions the person parameter estimates were found to be stochastically dependent on θ given ξ. As expected, there was a significant interaction effect between the latent variable ξ and θ with respect to $\hat{\xi}_{PCS}$. The contribution of the quadratic term ξ^2 is relatively small. The regression coefficient of the conditional regression of the estimator $\hat{\xi}_{PCS}$ on its estimand ξ is moderated by the latent response propensity θ. Recall that the person parameter estimates were increasingly positively biased by PCS the lower the latent ability is (see Figures 4.5 and 4.8). However, this is only valid if $Kor(\xi, \theta) > 0$, since the average proportion of missing data and, therefore, the bias due to PCS increases with decreasing ability. Accordingly, the slope of the tangent to the regression curve $\hat{\xi}_{PCS}$ on ξ decreases with lower values of ξ. In contrast to IAS, the latent variable constructed in a measurement model with PCS of missing responses is not a simple linear combination of the latent ability and the latent response propensity. Rather, the regression $E(\hat{\xi}_{PCS} \mid \xi, \theta$ is a nonlinear function of $(\xi, \theta))$. However, ξ_{PCS} depends on both ξ and θ and is, therefore, not a pure measure of the latent ability of substantial interest.

Table 4.5: Regression Coefficients, $t-$ and $p-$ Values of the Multiple Regression of ML Person Parameter Estimates (PCS) on the true values of θ and ξ (Data Example A).

Dependent Variable:	$\hat{\xi}_{ML}$ (PCS & 1PLM)				$\hat{\xi}_{ML}$ (PCS & 2PLM)			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
Intercept	-0.060	0.017	-3.463	< 0.001	-0.050	0.017	-2.898	0.004
ξ	0.831	0.020	41.831	< 0.001	0.715	0.020	35.876	< 0.001
θ	-0.002	0.020	-0.125	0.901	0.147	0.020	7.249	< 0.001
$\xi \cdot \theta$	0.291	0.044	6.700	< 0.001	0.224	0.044	5.141	< 0.001
ξ^2	-0.051	0.024	-2.164	0.031	-0.042	0.024	-1.761	0.078
θ^2	-0.088	0.025	-3.541	< 0.001	-0.053	0.025	-2.129	0.033

In summary, PCS suffers from the same theoretical inconsistencies as IAS. Here PCS was regarded as a data augmentation method, since the missing responses are simply replaced by numbers c_i. The implicit assumptions underlying this naive imputation model are as unrealistic as in the case of IAS. In fact, formally, IAS can be regarded as a special case of PSC with $c = 0$ for all items. However, Lord proposed to choose $c_i = 1/A_i$ with A_i as the number of response categories of item i. Here the considerations were confined

[9] Since $E(Y \mid X, Z) = E[E(Y \mid X, Z) \mid X]$. In a linear multiple regression that is, $E(\alpha_0 + \alpha_1 X + \alpha_2 Z + \alpha_3 XZ \mid X) = \alpha_0 + \alpha_1 X + E(\alpha_2 Z \mid X) + \alpha_3 E(Z \mid X)X$. Let $E(Z \mid X) = \beta_0 + \beta_2 X$, then $E(Y \mid X, Z) = (\alpha_0 + \alpha_2\beta_0) + (\alpha_1 + \alpha_2\beta_1 + \alpha_3\beta_0)X + \alpha_3\beta_1 X^2$.

to dichotomous items with $c = 0.5$ for all items $i = 1, \ldots, I$. The rationale given by Lord is intuitive at first glance. Each person u has a positive probability to solve an omitted or not-reached item $P(Y_i = 1 \,|\, U = u)$. Unfortunately, he assumes further that item responses to missing items would result from pure guessing. Formally, that means that the probability to solve an omitted item is a constant, which is always stochastically independent of any random variable such as U or the latent ability ξ. The implicit assumption of stochastic independence $Y_i \perp \xi \,|\, D_i = 0$ is ignored in the estimation equation of the model parameters. Fundamental properties of the ML estimates do not hold if PCS is used for missing responses. Accordingly, the item and person parameters were found to be systematically biased. Despite the similarity between PCS and IAS, the biases of item and person parameters estimates are quite different. The Rasch model leads to strongly biased item difficulty estimates and a marked variance reduction in the person parameter estimates. The item discriminations in the 2PLM are considerably underestimated depending on the proportion of missing data per item. The person parameter estimates are non-linearly related with the true latent variable ξ. The relation between the estimand ξ and the estimator $\hat{\xi}_{PCS}$ is moderated by the latent response propensity θ. Additionally, the estimated test information function, the standard error function, and the marginal reliability coefficient are heavily distorted. Due to the results PCS is not recommended as a strategy to deal with item non-responses, regardless of the underlying missing data mechanism. Even if the missing data mechanism is MCAR, PCS produces biased parameter estimates.

4.4 Nominal Response Model for Non-ignorable Missing Data

Bock (1972) proposed a latent-trait model for items with nominal response categories. This model is suited if $C > 2$ mutually exclusive, exhaustive and non-ordered categories exist (Baker & Kim, 2004). The nominal response model (NRM) rests upon the multivariate generalization of the logistic response function. It is strongly related to the multinomial logistic regression model (e. g. Agresti, 2002). The measurement model in the NRM is constituted by a set of I multinomial logistic regressions $P(Y_i = y_i \,|\, \xi)$. The item response category characteristic curve (IRCCC) are the conditional probability functions $P(Y_i = y_i \,|\, \xi)$ for each response category y_i of Y_i. Since the categories are exclusive and exhaustive it that follows $\sum_{y=0}^{C-1} P(Y = y \,|\, \xi) = 1$, implying that a response can and must occur in only one of the C response categories. Furthermore, local stochastic independence is assumed in the NRM. Bock proposed to apply the model to multiple choice items. For ex-

ample, Thissen (1976) applied the NRM to Raven's colored progressive matrices. When he analyzed the test information function he found that incorrect responses are informative with respect to the underlying trait of interest. As a result lower standard errors were obtained, especially in lower ranges of ξ. A missing response is also an incorrect answer but potentially informative with respect to a person?s ability. Recall that nonignorable missing data are also called informative missing responses. That is, the missing pattern has information that needs to be appropriately included in parameter estimation to ensure unbiased parameter estimates. The idea of using the NRM for item nonresponses is to consider a missing response as an additional incorrect response alternative. Hence, there are two distinct response behaviors resulting in an incorrect response - answering item i incorrectly, or failing to respond to item i. Both are assumed to have information with respect to the latent variable of interest. Against this background, it seems reasonable to apply the NRM to dichotomous data with non-ignorable missing data. Moustaki and O'Muircheartaigh (2000) suggested this approach first. To apply the NRM the manifest items Y_i are replaced by a new trivariate random variable R_i with a different domain and a different distribution. Formally, R_i is defined as a random variable R_i: $\Omega \rightarrow \Omega_{R_i}$ with $\Omega_{R_i} = \{0, 1, 2\}$. The values of R_i exclusively depend on Y_i and D_i. Hence, $R_i = f(Y_i, D_i)$. The function $f(Y_i, D_i)$ is given by the assignment rule

$$R_i = \begin{cases} 0, \text{if } Y_i = 0 \text{ and } D_i = 1 \\ 1, \text{if } Y_i = 1 \text{ and } D_i = 1 \\ 2, \text{if } D_i = 0. \end{cases} \tag{4.31}$$

Note that the categories of R_i do not have an inherent rank order. Prior to the examination of the applicability of the NRM to model item nonresponses, important properties of the model will be reviewed.

Initially, the NRM is considered without missingness. Accordingly, let Y_i be the manifest items with $C > 2$ unordered nominal response categories . The model equation of the NRM is then

$$P(Y_i = c \,|\, \xi) = \frac{\exp\left(\alpha_{i0c}^{(NR)} + \alpha_{i1c}^{(NR)} \cdot \xi\right)}{\sum_{h=0}^{C-1} \exp\left(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi\right)} \tag{4.32}$$

In comparison to two-parametric models for ordered categorical response variables such as the Graded Response Model (GRM; Samejima, 1969) or the Generalized Partial Credit Model (GPCM; Muraki, 1992), there are two category specific parameters - the inter-

cept $\alpha_{i0y}^{(NR)}$ and the discrimination parameter $\alpha_{i1y}^{(NR)}$ [10]. Restrictions are needed in order to identify the model. The interpretation of the model parameters depends on the chosen identification. Two different restrictions to identify the NRM are discussed in the literature (e. g. de Ayala, 2009). First, the sums $\sum_{c=0}^{C-1} \alpha_{i0c}^{(NR)}$ and $\sum_{c=0}^{C-1} \alpha_{i1c}^{(NR)}$ can be fixed to zero. Second, the parameters $\alpha_{i0c}^{(NR)}$ and $\alpha_{i1c}^{(NR)}$ with respect to one response category c of C needs to be fixed; typically $\alpha_{i0c}^{(NR)} = \alpha_{i1c}^{(NR)} = 0$. In this case, the category c serves as a reference category. This model is also called the baseline-category multinomial logit model (Agresti, 2002). For example, if we choose $\alpha_{i00}^{(NR)} = \alpha_{i10}^{(NR)} = 0$, then $Y_i = 0$ is the reference category with the conditional category probability

$$\begin{aligned} P(Y_i = 0 \,|\, \xi) &= \frac{\exp\left(\alpha_{i00}^{(NR)} + \alpha_{i10}^{(NR)} \cdot \xi\right)}{\exp\left(\alpha_{i00}^{(NR)} + \alpha_{i10}^{(NR)} \cdot \xi\right) + \sum_{h=1}^{C-1} \exp\left(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi\right)} \\ &= \frac{1}{1 + \sum_{h=1}^{C-1} \exp\left(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi\right)}. \end{aligned} \tag{4.33}$$

For a trivariate nominal response variable Y_i, the conditional category probabilities given ξ for the remaining two response categories are

$$P(Y_i = 1 \,|\, \xi) = \frac{\exp\left(\alpha_{i01}^{(NR)} + \alpha_{i11}^{(NR)} \cdot \xi\right)}{1 + \sum_{h=1}^{C-1} \exp\left(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi\right)} \tag{4.34}$$

$$P(Y_i = 2 \,|\, \xi) = \frac{\exp\left(\alpha_{i02}^{(NR)} + \alpha_{i12}^{(NR)} \cdot \xi\right)}{1 + \sum_{h=1}^{C-1} \exp\left(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi\right)}. \tag{4.35}$$

If a researcher decides to choose the NRM instead of the 1PLM or 2PLM to account for item nonresponses, the question is how the item parameters of these models are related theoretically. Therefore, can the item parameters of the 1PLM or 2PLM be estimated using the NRM? In the NRM for missing responses in dichotomous items, the manifest items Y_i in the measurement model are replaced by R_i. Although the variables Y_i and R_i are different, the item parameters of the 2PLM are preserved in the NRM if $R_i = 0$ is the reference category in the baseline-category multinomial logit model. In this case the odds

[10] In GRM and GPCM the item discrimination is constant across the response categories of an item Y_i.

ratio of the response categories $R_i = c$ ($c \in \{1, 2\}$) and $R_i = 0$ is

$$\frac{P(R_i = c \,|\, \xi)}{P(R_i = 0 \,|\, \xi)} = \frac{\exp\left(\alpha_{i0c}^{(NR)} + \alpha_{i1c}^{(NR)} \cdot \xi\right)}{\frac{1 + \sum_{h=1}^{C-1} \exp\left(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi\right)}{1 + \sum_{h=1}^{C-1} \exp(\alpha_{i0h}^{(NR)} + \alpha_{i1h}^{(NR)} \cdot \xi)}} \tag{4.36}$$

$$= \exp\left(\alpha_{i0y}^{(NR)} + \alpha_{i1y}^{(NR)} \cdot \xi\right). \tag{4.37}$$

The respective logit is

$$\ln\left(\frac{P(R_i = c \,|\, \xi)}{P(R_i = 0 \,|\, \xi)}\right) = \alpha_{i0c}^{(NR)} + \alpha_{i1c}^{(NR)} \cdot \xi. \tag{4.38}$$

For the case of dichotomous items Y_i and, therefore, trivariate manifest variables R_i, two non-redundant logits can be considered. The third logit is simply a function of the remaining logits. If $R_i = 0$ was chosen as the reference category, then the logarithm of the two odds $P(R_i = 1 \,|\, \xi)/P(R_i = 0 \,|\, \xi)$ and $P(R_i = 2 \,|\, \xi)/P(R_i = 0 \,|\, \xi)$ can be written as

$$\ln\left(\frac{P(R_i = 1 \,|\, \xi)}{P(R_i = 0 \,|\, \xi)}\right) = \alpha_{i01}^{(NR)} + \alpha_{i11}^{(NR)} \cdot \xi \tag{4.39}$$

$$\ln\left(\frac{P(R_i = 2 \,|\, \xi)}{P(R_i = 0 \,|\, \xi)}\right) = \alpha_{i02}^{(NR)} + \alpha_{i12}^{(NR)} \cdot \xi. \tag{4.40}$$

From the assignment rule used to construct R_i (see Equation 4.31) follows that $Y_i = R_i \,|\, D_i = 1$. Hence,

$$\ln\left(\frac{P(R_i = 1 \,|\, \xi)}{P(R_i = 0 \,|\, \xi)}\right) = \ln\left(\frac{P(Y_i = 1 \,|\, \xi, D_i = 1)}{P(Y_i = 0 \,|\, \xi, D_i = 1)}\right). \tag{4.41}$$

If $Y_i \perp D_i \,|\, \xi$ holds true (no DIF exist with respect to D_i) then

$$\ln\left(\frac{P(R_i = 1 \,|\, \xi)}{P(R_i = 0 \,|\, \xi)}\right) = \ln\left(\frac{P(Y_i = 1 \,|\, \xi)}{P(Y_i = 0 \,|\, \xi)}\right) \tag{4.42}$$

$$= \alpha_{i11}^{(NR)} \cdot \left(\xi - \beta_{i1}^{(NR)}\right), \tag{4.43}$$

with $\beta_{i1}^{(NR)} = -\alpha_{i01}^{(NR)}/\alpha_{i11}^{(NR)}$. The logit in Equation 4.42 is the same as in the Birnbaum model with the item difficulty $\beta_i = \beta_{i1}^{(NR)}$ and the item discrimination $\alpha_i = \alpha_{i01}^{(NR)}$. Hence, the item parameters of the 2PLM should be estimable using the NRM. In fact, if the latent variable ξ is known, then the basline-category logit model can be used to obtain unbiased

item parameters. In this case, the NRM is equivalent to a set of I multinomial logistic regression models with R_i as dependent variables and ξ taken as manifest predictors. Figure 4.10 and table 4.6 shows the item parameter estimates of Data Example A. Although

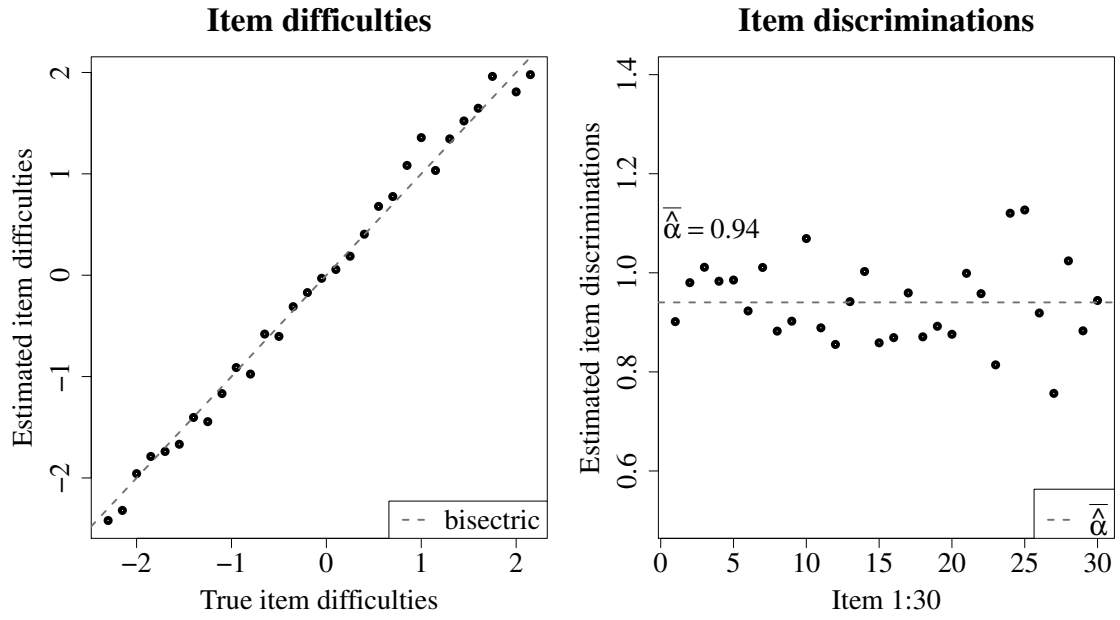

Figure 4.10: Item parameters of Data Example A estimated by multinomial logistic regression models with known values of ξ.

the item discriminations are, on average, slightly underestimated ($\bar{\hat{\alpha}} = 0.940$, $t = -3.870$, $df = 29$, $p = 0.001$), the multinomial regression models recover the true item parameters reasonably well. However, in real applications the underlying variable ξ is unknown and needs to be inferred from the realized data as well.

Person parameter estimation in the NRM for item nonresponses Although the NRM for non-ignorable missing data is different from naive imputation methods such as IAS or PCS, there are some similarities between the methods. The NRM is also estimated based on a data matrix $\boldsymbol{R} = \boldsymbol{r}$ that is free of missing data. The variables R_i are functions $f(Y_i, D_i)$ of two variables - the items Y_i and the response indicators D_i. As in the case of IAS and PCS, two pieces of information are combined in the manifest variables that constitute the measurement model of the latent variable: (a) the performance on item i, and (b) the information about the willingness or ability to complete the item, regardless of whether correctly or incorrectly. Given that the latent response propensity θ and the latent ability ξ exist with $\theta \neq \xi$ and Y_i is stochastically dependent from ξ and D_i is stochastically

dependent from θ, then R_i is stochastically dependent from both θ and ξ. If $\xi = f(U)$ and $\theta = f(U)$, then it follows from the construction of R_i that

$$\begin{aligned} P(R_i = 0 \,|\, U) &= P(Y_i = 0 \cap D_i = 1 \,|\, U) && (4.44) \\ &= P(Y_i = 0 \,|\, U, D_i = 1) P(D_i = 1 \,|\, U) \\ &= P(Y_i = 0 \,|\, \xi, D_i = 1) P(D_i = 1 \,|\, \theta) \\ P(R_i = 1 \,|\, U) &= P(Y_i = 1 \,|\, U, D_i = 1) && (4.45) \\ &= P(Y_i = 1 \,|\, U, D_i = 1) P(D_i = 1 \,|\, U) \\ &= P(Y_i = 1 \,|\, \xi, D_i = 1) P(D_i = 1 \,|\, \theta) \\ P(R_i = 2 \,|\, U) &= P(D_i = 0 \,|\, U) && (4.46) \\ &= P(D_i = 0 \,|\, \theta) \end{aligned}$$

Note that if conditional independence $Y_i \perp D_i \,|\, \xi$ holds true, then $P(Y_i = y \,|\, \xi, D_i = 1) = P(Y_i = y \,|\, \xi)$. If a parametric IRT model is valid, then the conditional probabilities $P(Y_i = y \,|\, \xi)$ are given by the respective measurement model equations. The derivations show that the manifest variables R_i that constitute the measurement model in the NRM for item nonresponses depends on both - the latent ability and the latent response propensity. This can also be demonstrated empirically utilizing Data Example A. Here the underlying realized values of the latent response propensity θ and the ability ξ are known. Multinomial logistic regressions $P(R_i \,|\, \theta)$ and $P(R_i \,|\, \xi)$ were used with *Nagelkerke's* pseudo -R^2 (Hu, Shao, & Palta, 2006; Nagelkerke, 1991) as a measure of the effect size of the stochastic dependency between R_i and θ or ξ, respectively. Table 4.6 shows the parameter estimates and *Nagelkerke's* R^2 for each variable R_i. On average, Nagelkerke's R^2 of the multinomial logistic regressions R_i on ξ was 0.175, and 0.196 of the regressions R_i on θ. The results confirm that the variables R_i are stochastically dependent from both, the latent ability ξ and the latent response propensity θ. Why is this important? Using the NRM for nonignorable missing data with R_i as manifest variables in the measurement model, only one single latent variable is constructed. Does R_i indicate the latent ability ξ or the latent response propensity θ, or even both? In other words, the meaning of the latent variable ξ_{NR} constructed in the NRM is questionable. If the person variable is not equal to ξ, then the item parameter estimates are also potentially different from the item parameters of the 2PLM. To emulate the real situation, the NRM for item nonresponses was applied to Data Example A. Both - item and person parameters - were estimated based on the manifest variables R_i. MULTILOG 7 (D. M. Thissen, Chen, & Bock, 2003) was used

Table 4.6: Parameter Estimates and *Nagelkerke's* R^2 for Multinomial Logistic Regressions $P(R_i = r_i \,|\, \xi)$ and $P(R_i = r_i \,|\, \theta)$ (Data Example A).

	Estimates of $P(R_i = r_i \,|\, \xi)$					Estimates of $P(R_i = r_i \,|\, \theta)$				
R_i	$\hat{\beta}_{i1}^{(NR)}$	$\hat{\beta}_{i2}^{(NR)}$	$\hat{\alpha}_{i11}^{(NR)}$	$\hat{\alpha}_{i12}^{(NR)}$	$R^2_{Y_i\|\xi}$ [1]	$\hat{\beta}_{i1}^{(NR)}$	$\hat{\beta}_{i2}^{(NR)}$	$\hat{\alpha}_{i11}^{(NR)}$	$\hat{\alpha}_{i12}^{(NR)}$	$R^2_{Y_i\|\theta}$ [1]
R_1	-2.420	-2.417	0.901	-0.054	0.152	-3.243	-0.805	0.626	-0.580	0.155
R_2	-2.320	-0.992	0.980	0.127	0.151	-2.783	-0.521	0.764	-0.326	0.155
R_3	-1.958	12.604	1.011	-0.016	0.188	-2.497	-0.095	0.734	-0.420	0.178
R_4	-1.788	2.915	0.983	-0.057	0.192	-1.959	-0.134	0.848	-0.425	0.214
R_5	-1.739	-14.337	0.985	0.073	0.181	-1.945	3.880	0.829	-0.234	0.199
R_6	-1.668	2.348	0.923	-0.128	0.195	-2.075	0.171	0.692	-0.550	0.209
R_7	-1.403	8.927	1.011	-0.069	0.215	-1.855	0.730	0.700	-0.562	0.229
R_8	-1.445	1.863	0.882	-0.134	0.187	-1.820	0.225	0.658	-0.473	0.187
R_9	-1.168	0.432	0.902	-0.097	0.185	-1.362	-0.253	0.722	-0.434	0.191
R_{10}	-0.911	0.726	1.069	-0.072	0.227	-1.116	-0.293	0.786	-0.505	0.221
R_{11}	-0.975	4.547	0.889	-0.156	0.196	-1.203	0.973	0.663	-0.596	0.231
R_{12}	-0.581	-0.002	0.855	-0.244	0.194	-0.700	-0.202	0.653	-0.596	0.207
R_{13}	-0.604	0.108	0.942	-0.177	0.207	-0.731	-0.221	0.705	-0.566	0.213
R_{14}	-0.310	1.754	1.003	-0.248	0.232	-0.386	0.463	0.719	-0.697	0.254
R_{15}	-0.172	-2.134	0.859	-0.243	0.185	-0.173	-0.996	0.648	-0.715	0.207
R_{16}	-0.031	0.170	0.869	-0.245	0.187	0.012	-0.118	0.691	-0.661	0.224
R_{17}	0.057	3.825	0.959	-0.317	0.210	0.098	1.777	0.780	-0.677	0.250
R_{18}	0.187	1.318	0.871	-0.421	0.216	0.226	0.693	0.655	-0.752	0.243
R_{19}	0.404	4.108	0.892	-0.348	0.177	0.504	2.045	0.651	-0.715	0.213
R_{20}	0.679	3.791	0.876	-0.360	0.163	0.816	2.074	0.767	-0.669	0.210
R_{21}	0.776	3.073	0.999	-0.420	0.196	0.932	1.758	0.790	-0.754	0.232
R_{22}	1.083	2.882	0.958	-0.525	0.189	1.242	2.052	0.778	-0.761	0.212
R_{23}	1.357	5.195	0.814	-0.430	0.111	1.577	3.504	0.686	-0.657	0.139
R_{24}	1.033	7.240	1.120	-0.230	0.138	1.301	2.878	0.851	-0.594	0.169
R_{25}	1.345	4.426	1.127	-0.385	0.149	1.570	2.700	0.913	-0.651	0.174
R_{26}	1.522	2.525	0.919	-0.455	0.153	1.929	1.516	0.679	-0.792	0.199
R_{27}	1.647	4.773	0.757	-0.443	0.102	2.167	2.834	0.546	-0.786	0.150
R_{28}	1.961	3.111	1.024	-0.578	0.146	2.155	2.170	0.946	-0.873	0.201
R_{29}	1.808	4.052	0.883	-0.533	0.123	2.359	2.956	0.592	-0.759	0.141
R_{30}	1.979	4.309	0.944	-0.494	0.111	2.380	2.854	0.752	-0.782	0.152

[1] Nagelkerke's Pseudo-R^2

to estimate model parameters. Item parameters were obtained by MML estimation with non-adaptive quadrature. 19 quadrature points were chosen. The latent variable was fixed to $E(\xi_{NR}) = 0$ and $Var(\xi_{NR}) = 1$. The person parameters were estimated in a second step using the item parameter estimates as fixed. Figure 4.11 shows the ML person parameter estimates $\hat{\xi}_{NR}$ compared to the true values of ξ and θ underlying Data Example A. It can be seen that the correlation between $\hat{\xi}_{NR}$ and the latent response propensity is even higher than the correlation of the latent ability and the ML person parameter estimates resulting from the NRM. Additionally, the partial correlations $r(\theta, \hat{\xi}_{NR}.\xi) = 0.685$ ($t = 42.026$, $df = 1998$, $p < 0.001$) and $r(\xi, \hat{\xi}_{NR}.\theta) = 0.319$ ($t = 15.025$, $df = 1998$, $p < 0.001$) deviate

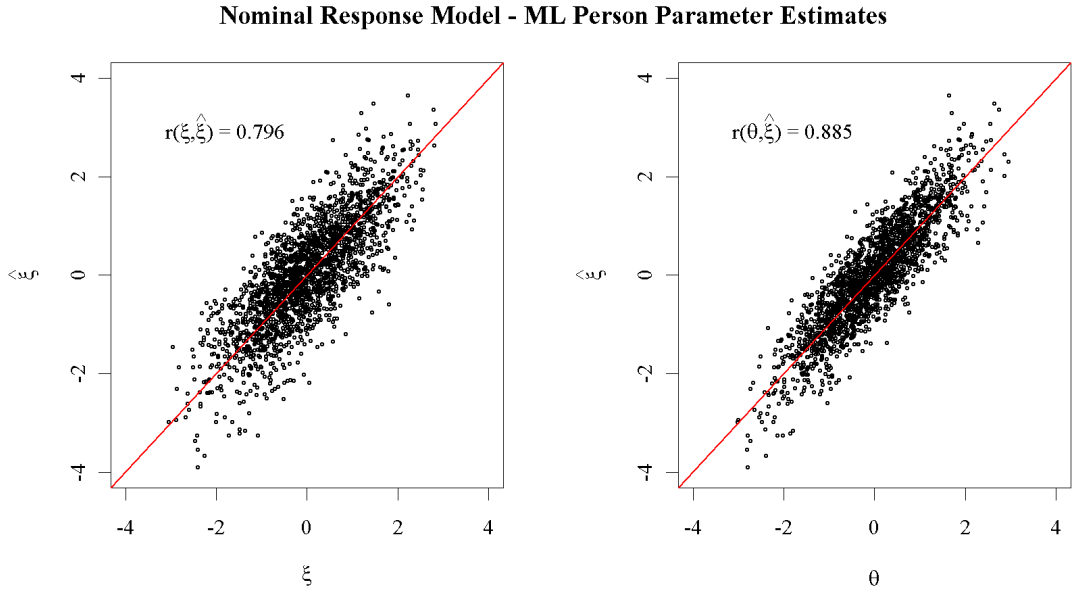

Figure 4.11: Relationship between ML person parameter estimates $\hat{\xi}_{ML}$ of the NRM and values of the latent ability (left) and the latent response propensity (right). The red lines indicate the bisectric.

significantly from zero. Furthermore, the parameters of a multiple regression $E(\hat{\xi}_{NR}\,|\,\xi,\theta)$ were estimated. Additionally, the determination coefficient was $R^2_{\hat{\xi}_{NR}\,|\,\xi,\theta} = 0.805$. This is significantly higher than the proportions of explained variances in the two simple regressions $E(\hat{\xi}_{NR}\,|\,\xi)$ with $R^2_{\hat{\xi}_{NR}\,|\,\xi} = 0.633$ ($R_{diff.} = 0.172$, $F = 883.04$, $df_1 = 2$, $df_2 = 1996$, $p < 0.001$) and $E(\hat{\xi}_{NR}\,|\,\theta)$ with $R^2_{\hat{\xi}_{NR}\,|\,\theta} = 0.783$ ($R_{diff.} = 0.022$, $df_1 = 2$, $df_2 = 1996$, $p < 0.001$). As Table 4.7 shows, the partial standardized regression coefficients of both latent variables are significantly different from zero. As expected, the latent variable constructed in the NRM based on the manifest variables R_i is also a linear combination of the latent ability and the latent response propensity. As in the case of IAS, the confusion of two different pieces of information given by the variables Y_i and D_i is reflected in the latent variable in the NRM. Due to the substantial correlation $r(\xi,\hat{\xi}_{NR}) = 0.796$, it might be tempting to conclude that the NRM recovers the ability ξ well. However, such a high correlation cannot be generally expected. The missing data mechanism is essential for the correlation $Cor(\xi,\xi_{NR})$. Since ξ_{NR} is a linear combination of ξ and θ, it is expected that the correlation $Cor(\xi_{NR},\xi)$ decreases, the lower the correlation $Cor(\xi,\theta)$ is. Additionally, it is hypothesized that the overall proportion of missing data affects $Cor(\xi_{NR},\xi)$ and, therefore, the meaning of $\hat{\xi}_{NR}$. The higher the proportion of missing data is, the less in-

Table 4.7: Estimated Regression Coefficients, Standard Errors (*SE*), *t*- and *p*-values for the Multiple Regression $E(\hat{\xi}_{NR} \mid \xi, \theta)$.

Independent variable	Coeff.[1]	*SE*	*t*	*p*
ξ	0.269	0.018	15.016	< 0.001
θ	0.765	0.018	41.978	< 0.001
$\theta * \xi$	-0.008	0.009	-0.915	0.360

[1] Standardized partial regression coefficients.

formation about the performance on the single items and the achievement at the complete test from the response behavior result. Hence, the variables R_i are less informative with respect to the latent ability ξ, which should be reflected by lower correlations $Cor(\xi_{NR}, \xi)$. Conversely, it is expected that with rising probabilities of the occurrence of missing data the variables R_i are more informative with respect to the latent response propensity θ, resulting in higher correlations $Cor(\xi_{NR}, \theta)$. Of course, additional factors, that were not considered here, might influence the construction of ξ_{NR} and the correlation $Cor(\xi_{NR}, \xi)$. In a short simulation study with only two factors - the correlation $Cor(\xi, \theta)$ and the overall proportion of missing data - it could be shown that $Cor(\xi_{NR}, \xi)$ varies considerably. The correlation $Cor(\xi, \theta)$ was chosen to be 0, 0.2, 0.5, 0.8, and 1. The overall proportions of missing data were approximately 10%, 20% and 50%. Note that a correlation of $Cor(\xi, \theta) = 0$ means the missing data mechanism is MCAR. In general, more difficult items had higher probabilities to be omitted. The measurement model consisted of 30 dichotomous items. The item parameters used for simulation were the same as in Data Example A (see Table 3.1). The sample size was $N = 2000$ for all simulated data examples. After the simulation of $\boldsymbol{Y}$ and $\boldsymbol{D}$ the manifest variables R_i were generated according to Equation 4.31. The resulting 15 data sets were analyzed using the NRM implemented in MULTILOG 7. The results are summarized in Table 4.8. Due to the small number of simulated data sets, no statistical analysis of the results was conducted. In general, the correlation $r(\xi, \hat{\xi}_{NR})$ was high given the proportion of missing data was small. The higher the proportion of missing data was, the higher $r(\xi, \hat{\xi}_{NR})$ was, depending on the correlation $Cor(\xi, \theta)$. For example, if the proportion of missing data was 30%, then $r(\xi, \hat{\xi}_{NR})$ ranges between 0.428 and 0.918, depending on $Cor(\xi, \theta)$. For the case of 50% proportion of missing data, $r(\xi, \hat{\xi}_{NR})$ dropped to 0.148 when $Cor(\xi, \theta) = 0$ but remained high ($r(\xi, \hat{\xi}_{NR}) = 0.906$) when $Cor(\xi, \theta) = 1$. Thus, there is an interaction effect between the overall proportion of missing data and $Cor(\xi, \theta)$ with respect to $Cor(\xi, \xi_{NR})$. On the other hand, the correlation $Cor(\theta, \xi_{NR})$ varied the most depending on $Cor(\xi, \theta)$ if the overall

Table 4.8: Correlations and Partial Correlations of ML person Parameter Estimates of the NRM and the True Values of ξ and θ Under Different Conditions.

% Missings	$Cor(\xi, \theta)$	$r(\xi, \hat{\xi}_{NR})$	$r(\theta, \hat{\xi}_{NR})$	$r(\xi, \hat{\xi}_{NR}.\theta)$	$r(\theta, \hat{\xi}_{NR}.\xi)$	$Rel(\hat{\xi}_{NR})$
10	0.0	0.842	0.153	0.864	0.385	0.805
10	0.2	0.830	0.461	0.843	0.517	0.828
10	0.5	0.874	0.637	0.831	0.467	0.847
10	0.8	0.903	0.831	0.700	0.384	0.860
10	1	0.921	0.921	/	/	0.869
20	0.0	0.428	0.718	0.624	0.799	0.803
20	0.2	0.566	0.738	0.657	0.787	0.818
20	0.5	0.733	0.787	0.661	0.732	0.852
20	0.8	0.853	0.869	0.550	0.611	0.871
20	1	0.918	0.918	/	/	0.885
50	0.0	0.148	0.862	0.321	0.874	0.839
50	0.2	0.356	0.850	0.344	0.848	0.835
50	0.5	0.589	0.865	0.355	0.814	0.849
50	0.8	0.768	0.888	0.239	0.717	0.863
50	1	0.906	0.906	/	/	0.871

proportion of missing data was small ($0.153 \leq r(\theta, \xi_{NR}) \leq 0.921$). Hence, there is also an interaction effect between the overall proportion of missing data and $Cor(\xi, \theta)$ with respect to $Cor(\theta, \xi_{NR})$. Generally, in all data examples $r(\theta, \xi_{NR}.\xi)$ deviated substantially from zero. This highlights that ξ_{NR} is indeed a linear combination of both underlying latent variables - ξ and θ - plus a stochastic component (residual).

Two results are of major importance. First, the more the correlation between the latent response propensity and the latent ability deviates from one, the more the correlation $Cor(\xi, \xi_{NR})$ decreases. Absurdly, the NRM yields the worst parameter recovery when the missing data $\boldsymbol{Y}$ is MCAR and yields the best parameter recovery if the missing data mechanism is NMAR with the latent response propensity as a linear function of ξ. In this particular situation the person parameter estimation is unbiased. Second, the marginal reliabilities $Rel(\hat{\xi}_{NR})$ of the 15 data sets estimated with MULTILOG 7 were almost equal across the simulated data sets regardless the correlations $r(\xi, \hat{\xi}_{NR})$. Neither the overall proportion of missing data nor the correlation $Cor(\xi, \theta)$ between the latent ability and the latent response propensity lowers the marginal reliability substantially. An applied researcher may be lulled into a false sense of security in view of such good reliability coefficients in a seemingly valid and useful measurement model. Unfortunately, the problem

can hardly be detected in real applications.

Item parameter estimation in the NRM for item nonresponses Finally, the coverage of the item parameters of the 2PLM is considered if the NRM is used. In the beginning of this section it was shown how the item parameters of the 2PLM and the NRM for item nonresponses are related theoretically. It was also illustrated that the item parameters could be estimated unbiasedly if the true values of the latent variable are known. In real applications this is typically not the case. Rather, the individual values of ξ need to be jointly estimated from the data with the item parameters. Since ξ and ξ_{NR} are most likely not equal in real applications, the item parameters of the NRM are expected to be different from those of the 2PLM as well. In Figure 4.12 the item parameter estimates $\hat{\beta}_{i1}^{NR}$ and the discriminations $\hat{\alpha}_{i11}^{NR}$ of three simulated data sets, with $Cor(\xi, \theta)$ equal to 0.2, 0.8 (Data Example A), and 1 are displayed in comparison to the true item parameters, β_i and α_i. The item parameter estimates of two out of the 16 data sets from Table 4.8 with an overall proportion of 50% missing data were used. If the NRM yields unbiased item parameter estimates, then all estimates $\hat{\alpha}_{i11}^{NR}$ should be close to one. Using a scatter plot, the estimated values $\hat{\beta}_{i1}^{NR}$ and true difficulties should lie close to the bisectric. The item difficulty estimates $\hat{\beta}_{i1}^{NR}$ were consistently overestimated in Data Example A. The mean bias was 0.832, which differs significantly from zero ($t = 12.312, df = 29, p < 0.001$). The mean of the estimated item discriminations was $\bar{\hat{\alpha}}_{11}^{NR} = 0.831$, which is significantly lower than one ($t = -3.175$, $df = 29$, $p = 0.004$). If the correlation $Cor(\xi, \theta) = 0.2$, then the item difficulties were, on average, even more overestimated, whereas the item discriminations were, on average, more negatively biased. Unbiased item parameter estimates were only found if $Cor(\xi, \theta) = 1$ (lower two graphs of Figure 4.8). In this case, ξ and ξ_{NR} are linear functions of each other. If the 2PLM and the NRM are identified in the same way, for example if $E(\xi) = E(\xi_{NR}) = 0$ and $Var(\xi) = Var(\xi_{NR}) = 1$, then $\xi = \xi_{NR}$. In this case, the estimates $\hat{\alpha}_{i11}^{NR}$ and $-\hat{\alpha}_{i01}^{(NR)}/\hat{\alpha}_{i11}^{(NR)}$ of the NRM are unbiased estimates of α_i and β_i of the 2PLM. However, the equality $\theta = f(\xi)$, with $f(.)$ a linear function, has some important implications, because in this case the response indicators D_i could simply be used as additional manifest indicators of the latent ability ξ. Put simply, the variables D_i can be used as *additional items* in a joint measurement model based on $(\boldsymbol{Y}, \boldsymbol{D})$ with the assumption of local stochastic independence $Y_i \perp (\boldsymbol{D}, \boldsymbol{Y}^{-i}) | \xi$. The item- and person parameter estimates of both, the NRM and the unidimensional IRT model based on Y_i and D_i, should recover the true parameters equally well. Indeed, in the simulated data example with $\xi = \theta$ the correlation between the ML estimates of ξ obtained by the uni-

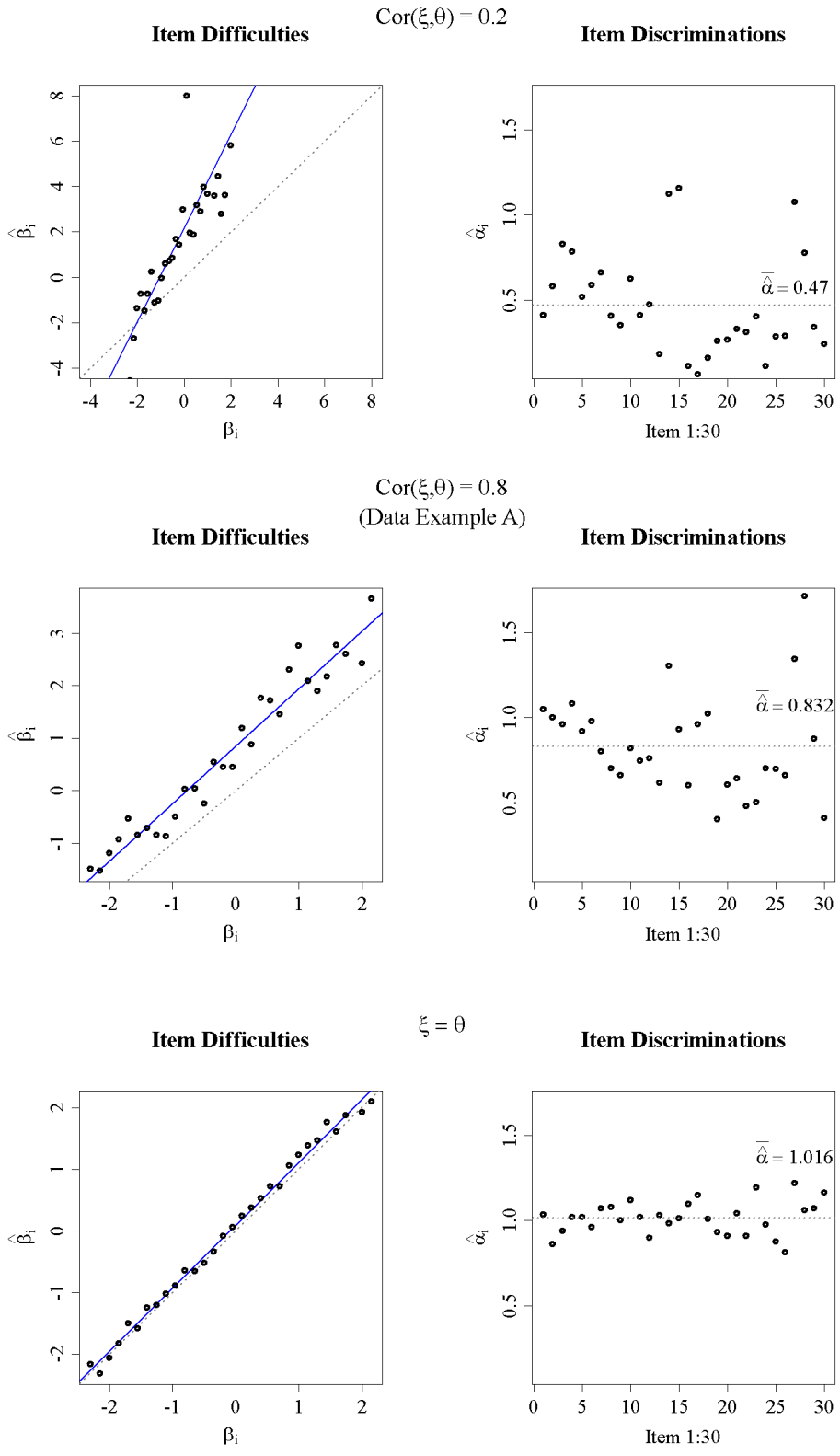

Figure 4.12: True and estimated item difficulties and discrimination parameters using the NRM in three different conditions: $Cor(\xi, \theta) = 0.2$, 0.8, and 1. The grey lines indicate bisectric lines (left column) or the means $\bar{\hat{\alpha}}$ (right column).

dimensional IRT model with 60 items $Y_1, \ldots, Y_I, D_1, \ldots, D_I$) was $r(\hat{\xi}_{ML}, \xi) = 0.967$. This value was even higher than $r(\hat{\xi}_{NR}, \xi) = 0.935$ using the NRM and $r(\hat{\xi}_{ML}, \xi) = 0.910$ using the complete data $\boldsymbol{Y} = \boldsymbol{y}$ for person parameter estimation.

Summary It could be shown that IAS, PCS, and the NRM for item nonresponses have a common problem. The manifest variables used in the measurement model are different form the original variables Y_i. If both - item and person parameters - are unknown, then parameter estimation will most likely be biased. Strictly speaking, the construction of the latent variable is affected. The replacement of the manifest indicators Y_i by Y_i^* and R_i or the imputation of c in PCS results in substantially different models with different parameters that have a different meaning. The item and person parameter estimates are biased in the sense that they systematically differ from the parameters *aimed* to be estimated. As in the case of IAS, the latent variable ξ_{NR} in the NRM for item nonresponses is a linear combination of the latent ability and the latent response propensity. Unbiased parameter estimates of the 2PLM can only be obtained by the NRM if the latent response propensity is a linear function of ξ. In this case all item response propensities are functions of the latent ability. However, in this case a unidimensional IRT model including both, the items Y_i and the corresponding response indicators D_i, could be used alternatively. Neither the existence of a latent response propensity nor the correlation between the latent variables ξ and θ can be examined in the NRM. Thus, the question of whether the NRM is appropriate to account for nonignorable missing data is directly related to the question of dimensionality in common measurement model based on $\boldsymbol{Y}$ and $\boldsymbol{D}$. Interestingly, multidimensional IRT models including a model of a latent response propensity have been proposed as a model based approach for nonignorable missing data. Such models do not require that $Cor(\xi, \theta) = 0$. Furthermore, the flexibility of these models allow for multidimensional latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. In the following sections, MIRT models for missing responses that are NMAR will be of major interest.

4.5 IRT Model Based Methods

In the previous section, it could be shown that naive imputation methods such as IAS and PCS cannot be recommended to handle item nonresponses in most applications. The NRM for item nonresponses can be regarded as a model-based method that yields unbiased parameter estimates of the 2PLM if strong assumptions hold true. In this section, less restrictive model based approaches for missing data in IRT measurement models will be introduced and further developed. The major focus is put on models for nonignorable item nonresponses. Of course, methods for missing responses in measurement models that are MCAR or MAR are no less important but have been addressed in many publications and are already implemented in many mainstream software programs such as M*plus* 6 (Muthén & Muthén, 1998 - 2010) or LISREL 8 (Jöreskog & Sörbom, 1997, 2006). FIML estimation and parameter estimation by means of EM algorithms are the most popular among the model based approaches. Furthermore, multiple imputation of ignorable item nonresponses in dichotomous items has been found to work very well even if the proportion of missing data is substantial (Van Buuren, 2007, 2010). Methods to handle nonignorable missing data, however, are much less commonly used.

In this chapter, ML estimation with missing data will be examined first considering the different missing data mechanisms as defined in Section 2.1. Based on these general derivations, appropriate IRT models for ignorable missing data will be briefly reviewed. Then models for nonignorable item nonresponses in dichotomous items are introduced and further developed. Outside the field of measurement, two major classes of models for nonignorable missing data have been proposed: (a) selection models (SLM; Heckman, 1976, 1979; Little, 2008; Winship & Mare, 1992) and (b) pattern mixture models (PMM; Heckman, 1976, 1979; Little, 2008; Winship & Mare, 1992). The IRT models examined here will be derived from these models. The underlying assumptions are made explicitly. Finally, it will be shown how unbiasedly esimated IRT item and person parameter estimates can be used to obtain corrected item means and sample distributions of the sum score.

4.5.1 ML Estimation in IRT Models With Missing Data

In his seminal paper, Rubin (1976) explicated the conditions required to hold that sample based inference in presence of missing data is valid. He considered inference based on ML estimation as well Bayesian estimation methods. Here, the considerations will be confined to ML estimation with missing data. In Section 4.2 ML estimation in the case

of complete data was already introduced. Now, it will be extended to cases of incomplete data following Rubin (1976), Little & Rubin (2002), Schafer (1997), and Mislevy & Wu (1996).

As shown in Section 4.2, the ML estimate $\hat{\boldsymbol{\iota}}_{ML}$ is found by maximizing the probability $g(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\iota})$ of the observed data $\boldsymbol{Y} = \boldsymbol{y}$ given the parameters $\boldsymbol{\iota}$. In application that means that the likelihood function $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$ (see Equation 4.1), which is any function that is proportional to $g(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\iota})$, is maximized. If there is any missing data mechanism, the observed data of $\boldsymbol{Y} = \boldsymbol{y}$ reduces to $\boldsymbol{Y}_{\boldsymbol{obs}} = \boldsymbol{y}_{\boldsymbol{obs}}$. Accordingly, ML estimation based on the observable data means to find the ML estimator by maximizing the likelihood function $\mathcal{L}(\boldsymbol{y}_{\boldsymbol{obs}}; \boldsymbol{\iota})$, which is proportional to $g(\boldsymbol{Y}_{\boldsymbol{obs}} = \boldsymbol{y}_{\boldsymbol{obs}}; \boldsymbol{\iota})$. As Rubin stated, sample-based inference is then conditional given the particular missing pattern $\boldsymbol{D} = \boldsymbol{d}$. The decisive question is, under which conditions sample-based ML estimates obtained from $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\iota})$ and $\mathcal{L}(\boldsymbol{y}_{\boldsymbol{obs}}; \boldsymbol{\iota})$ are asymptotically equal? If they are equal, then the following equation is implied:

$$E(\hat{\boldsymbol{\iota}}_{ML} \,|\, \boldsymbol{D} = \boldsymbol{d}) = E(\hat{\boldsymbol{\iota}}_{ML} \,|\, \boldsymbol{D} = \mathbf{1}) = \boldsymbol{\iota} \tag{4.47}$$

$\boldsymbol{D} = \mathbf{1}$ means there are no missing data. Hence, $E(\hat{\boldsymbol{\iota}}_{ML} \,|\, \boldsymbol{D} = \mathbf{1})$ is the expected value of the ML estimate in absence of missingness. If Equation 4.47 holds true, missing data leads to increased uncertainty due to the loss of information with respect to the estimand $\boldsymbol{\iota}$, reflected by larger standard errors. However, point estimates are asymptotically consistent and unbiased. The sufficient conditions required to ensure valid sample based inference in presence of missing data are also called ignorability conditions (Heitjan, 1994; Little & Rubin, 2002; Rubin, 1976). Prior to the introduction of the ignorability conditions, likelihood inference in presence of missing data is introduced.

Initially, the derivations are very general and not restricted to IRT measurement models. It will not be distinguished between (in)dependent variables or covariates. Hence, let $\boldsymbol{Y}$ be simply a $(N \times I)$ random matrix including the covariates with stochastically independent rows. If the respective random experiment is repeated N times, then the realized data matrix $\boldsymbol{y}$ results. If there is any missing data mechanism, then the observed missing pattern $\boldsymbol{d}$ is a realization of the $(N \times I)$ random matrix $\boldsymbol{D}$. The rows of $\boldsymbol{D}$ are also assumed to be stochastically independent. Hence, the missing patterns between the test takers are independent.

In order to study ML estimation with missing data, different likelihood functions are distinguished: (a) $\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ the full likelihood of the complete data, (b) $\mathcal{L}(\boldsymbol{y}_{\boldsymbol{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$

the full likelihood of the observed data, and (c) $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ denotes the observed data likelihood ignoring the missing data mechanism. The full likelihood $\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ refers to a joint model of $(\boldsymbol{Y}, \boldsymbol{D})$. Recall that $\boldsymbol{Y}$ and $\boldsymbol{D}$ are both random variables defined on the same probability space. The stochastic relationship between the response indicator variables Y_i of $\boldsymbol{Y}$ and D_i of $\boldsymbol{D}$ was used to define the missing data mechanisms w.r.t. Y_i. Similarly, the stochastic relationship between the partitions $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ of $\boldsymbol{Y}$ and $\boldsymbol{D}$ was used to define the missing data mechanism w.r.t. $\boldsymbol{Y}$ (see Section 2.1). The parameter vector $\boldsymbol{\iota}$ consists of the estimands of substantial interest. $\boldsymbol{\phi}$ is the parameter vector of the model of $\boldsymbol{D}$, for example, parameters of logistic regression of variables D_i on variables in $\boldsymbol{Y}$. Hence, the full likelihood functions refer to a joint model that includes the target model of interest and the model of missingness, which is typically not of researchers? interest. However, if the missing data mechanism is nonignorable, then the conditional distribution of the missing part $\boldsymbol{Y}_{mis}$ given the observable part $\boldsymbol{Y}_{obs}$ depends on $\boldsymbol{D}$. Missingness is then called informative with respect to the model of $\boldsymbol{Y}$ given by the parameter vector $\boldsymbol{\iota}$. Accordingly, $\boldsymbol{D}$ needs to be included in a joint model of $(\boldsymbol{Y}, \boldsymbol{D})$ to ensure unbiased parameter estimation of $\boldsymbol{\iota}$ if the the missing data mechanism is nonignorable. In contrast, if certain conditions - the so called ingorability conditions - hold true, then $\boldsymbol{D}$ does not need to be modelled jointly with $\boldsymbol{Y}$. $\boldsymbol{D}$ can be *ignored* and the observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ yields unbiased parameter estimates. Direct likelihood inference is based on ratios of likelihoods (Dempster, 1997; A. W. F. Edwards, 1972; Rubin, 1976). Two ignorability conditions need to hold true to ensure valid direct likelihood inference: (a) the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MCAR or MAR, and (b) the common parameter space $\Omega_{\boldsymbol{\iota},\boldsymbol{\phi}}$ can be written as $\Omega_{\boldsymbol{\iota}} \times \Omega_{\boldsymbol{\phi}}$. In a Bayesian sense, this implies that each $(\boldsymbol{\iota}, \boldsymbol{\phi}) \in \Omega_{\boldsymbol{\iota},\boldsymbol{\phi}}$ has a non-zero probability. In other words, this means that regardless of the dependency between the manifest variables $\boldsymbol{Y}$ and $\boldsymbol{D}$, which define the missing data mechanism, the range of defined values of $\boldsymbol{\iota}$ must not be restricted given particular values of $\boldsymbol{\phi}$ and vice versa. If these two conditions (a) and (b) hold true, then the missing data mechanism is called *ignorable* (Rubin, 1976). ML estimation resting upon $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ instead of $\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ is sufficient. This can be shown by considering the different likelihood functions introduced above and their relationships.

Relationship between the full likelihood and the likelihood ignoring the missing data mechanism The full complete data likelihood $\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) = \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ is proportional to the joint distribution $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$. Factorizing the

joint probability (Cramér, 1949; Cox & Wermuth, 1999) yields

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\phi}) \tag{4.48}$$

The full likelihood of the complete data can be considered theoretically but is never available in application since $\boldsymbol{y}_{mis}$ is always unobserved. Any applicable ML estimator needs to be based on the observable variables $\boldsymbol{Y}_{obs}$ and $\boldsymbol{D}$ with the respective full observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$, which is the integral of the right hand side of Equation 4.48

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto \int g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{Y}_{mis}; \boldsymbol{\iota}, \boldsymbol{\phi}) dY_{mis} \\ &\propto \int g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) dY_{mis} \end{aligned} \tag{4.49}$$

As Mislevy & Wu (1996) stated, under local stochastic independence $Y_i \perp Y_j \,|\, \boldsymbol{\xi}$ for all $i \neq j$ Equation 4.49 is

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto \int g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) dY_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) dY_{mis}. \end{aligned} \tag{4.50}$$

In both Equations 4.49 and 4.50 the integral is taken over all possible values $\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}$ under the given missing pattern $\boldsymbol{D} = \boldsymbol{d}$. If all variables Y_i are discrete, then a finite number of possible patterns $\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}$ in $\Omega_{Y_{mis}}$ exist with the probabilities $P(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota})$. The integrals in Equations 4.49 and 4.50 become a sum over the values in $\Omega_{Y_{mis}}$[11]. If the missing data mechanism is MCAR or MAR, then conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ holds true. In this case, the conditional distribution $g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\phi}) = g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi})$, implying

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) dY_{mis} \\ &\propto g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) dY_{mis} \\ &\propto g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \cdot 1. \end{aligned} \tag{4.51}$$

[11] For example, if all variables Y_i are dichotomous, then $\Omega_{Y_{mis}}$ consists of 2^Q possible pattern $\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}$, where $Q = \sum_{n=1}^{N} \sum_{i=1}^{I} D_{ni}$.

Hence, if the missing data mechanism is MCAR or MAR, then the integral in Equation 4.50 is equal to the integral $\int g(\boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}; \boldsymbol{\iota}) dY_{mis}$, which is always one. As a result, the likelihood given by Equation 4.51 factors into two pieces that can be maximized separately in order to find the estimates of $\boldsymbol{\iota}$ or $\boldsymbol{\phi}$. Hence, if only $\boldsymbol{\iota}$ is of interest, then it is sufficient to maximize the likelihood function $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$, which is proportional to $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota})$ (Rubin, 1976; Little & Rubin, 2002; Schafer, 1997; Mislevy & Wu, 1996).

In order to proof that direct likelihood inference is valid, the likelihoods are set to be equal to the respective probability density functions. Let $\boldsymbol{\iota}_1$ and $\boldsymbol{\iota}_2$ be two vectors defined in $\Omega_{\boldsymbol{\iota}}$. The likelihood ratio of the observed data likelhood ignoring missing data with respect to $\boldsymbol{\iota}_1$ and $\boldsymbol{\iota}_2$ is

$$\frac{\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota}_1)}{\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota}_2)} = \frac{g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_1)}{g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_2)}. \tag{4.52}$$

Given the missing data mechanism is ignorable, from Equation 4.51 follows that the likelihood ratio of the full likelihood of the observed data with respect to $\boldsymbol{\iota}_1$ and $\boldsymbol{\iota}_2$ is

$$\begin{aligned} \frac{\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}_1, \boldsymbol{\phi})}{\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}_2, \boldsymbol{\phi})} &= \frac{g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_1)}{g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_2)} \\ &= \frac{g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_1)}{g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}_2)} \\ &= \frac{\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota}_1)}{\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota}_2)} \end{aligned} \tag{4.53}$$

Hence, the likelihood ratios of the full likelihood of the observed data and the likelihood of the observed data ignoring missing data are equal. It is important to note that this equality does not necessarily follow, if the missing data mechanism is nonignorable. Direct likelihood inference is not generally valid in this case. A joint model of $(\boldsymbol{Y}, \boldsymbol{D})$ needs to be built with a joint estimation of $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$ even if the latter is not of substantial interest. In less technical terms, that means that the information of the missingness given by $\boldsymbol{D}$ with respect to $\boldsymbol{\iota}$ needs to be included to ensure unbiased parameter estimation and valid inference.

ML estimation with missing data in psychological and educational measurement In this work, the taxonomy for missing data introduced by Rubin was extended (see Section 2.2). Due to the differentiation between $\boldsymbol{Y}$ the manifest variables in the measurement model and $\boldsymbol{Z}$ a potentially multidimensional covariate, three different MAR conditions

can be distinguished. In the remainder of this section, the ML estimation in measurement models including covariates will be examined.

In the common missing data literature, it is not distinguished between covariates and independent or dependent variables. In fact, such a distinction is not necessary for general considerations and proofs as demonstrated previously. The following differentiation, however, is quite useful to derive appropriate models that allow for valid likelihood inference depending on the respective missing data mechanism. If $\boldsymbol{Z}$ is included in the model, then the likelihood function is proportional to the joint probability $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{Z} = z; \boldsymbol{\iota}, \boldsymbol{\phi})$, which can be factorized in many different ways. Similar to Equation 4.48, one possible factorization is

$$\begin{aligned} g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{Z} = z; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \\ &\quad \cdot g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{Z} = z; \boldsymbol{\phi}). \end{aligned} \tag{4.54}$$

The vector $\boldsymbol{\iota}$ contains not merely the item and person parameters of the measurement model but additional parameters that describe the stochastic dependency between $\boldsymbol{Y}$ and the covariate $\boldsymbol{Z}$, For example, regression coefficients and residual variances and covariances in a latent regression model $E(\boldsymbol{\xi} \,|\, \boldsymbol{Z})$. Since the person parameters are formally estimands included in $\boldsymbol{\iota}$, local stochastic independence $Y_i \perp Y_j \,|\, \boldsymbol{\xi}$ (for all $i \neq j$) implies

$$\begin{aligned} g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{Z} = z; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \\ &\quad \cdot g(\boldsymbol{Y}_{mis} = \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \\ &\quad \cdot g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{Z} = z; \boldsymbol{\phi}). \end{aligned} \tag{4.55}$$

If $\boldsymbol{Y} \perp \boldsymbol{Z} \,|\, \boldsymbol{\xi}$, meaning that no differential item functioning (DIF) with respect to $\boldsymbol{Z}$ exists, then Equation 4.55 can be further simplified to

$$\begin{aligned} g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{Z} = z; \boldsymbol{\iota}, \boldsymbol{\phi}) &= g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{Y}_{mis} = \boldsymbol{Y}_{mis}; \boldsymbol{\iota}) \\ &\quad \cdot g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{Z} = z; \boldsymbol{\phi}). \end{aligned} \tag{4.56}$$

The full likelihood $L(\boldsymbol{y}_{obs}, \boldsymbol{d}, z; \boldsymbol{\iota}, \boldsymbol{\phi})$ of the observed data can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, z; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto\ & g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \\ & \cdot \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) d\boldsymbol{Y}_{mis}. \end{aligned} \tag{4.57}$$

This equation refers to Equation 4.50, where no distinction was made between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. The observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ that ignores missing data is proportional to the joint distribution $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota})$. In order to answer the question whether $\boldsymbol{Z}$ needs to be modelled jointly with $\boldsymbol{Y}$ to ensure valid item and person parameter estimation, the likelihood $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ of the observed item responses is considered, which is proportional to $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota})$. In the following, ML estimation in IRT models with covariates will be examined for each missing data mechanism separately starting from Equation 4.57 to answer the question under which conditions unbiased item and person parameter estimation and valid likelihood inference is implied.

ML estimation in IRT models if the missing mechanism w.r.t. $\boldsymbol{Y}$ is MCAR If the missing data mechanism w. r. t $\boldsymbol{Y}$ is missing completely at random, defined by stochastic independence $\boldsymbol{D} \perp (\boldsymbol{Y}, \boldsymbol{Z})$ (see Equation 2.15), then the full likelihood of the observed data is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, z; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}). \qquad (4.58)
\end{aligned}$$

Hence, the observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ ignoring missingness is sufficient for unbiased ML estimation and valid direct likelihood inference. Furthermore, it can be shown that the covariate $\boldsymbol{Z}$ needs not to be modelled jointly with $\boldsymbol{Y}$ for estimation of parameters of the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ since

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\phi}). \qquad (4.59)
\end{aligned}$$

The observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ of the observed item responses gives unbiased item and person parameter estimates.

ML estimation in IRT models if the missing mechanism is MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$ The missing data mechanism w.r.t. $\boldsymbol{Y}$ has been defined to be missing at random given $(\boldsymbol{Y}, \boldsymbol{Z})$ if conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, (\boldsymbol{Y}_{obs}, \boldsymbol{Z})$ applies although conditional

stochastic dependence $\boldsymbol{D} \not\perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Z}$ and $\boldsymbol{D} \not\perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ holds true (see Equations 2.17). The full likelihood of the observed data is

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, z; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\phi}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\phi}) \end{aligned} \tag{4.60}$$

Again, the observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ ignoring the missing data mechanism is sufficient for valid direct likelihood inference and unbiased ML estimation. However, in contrast to MCAR, the covariate $\boldsymbol{Z}$ has to be included in a joint model of $(\boldsymbol{Y}, \boldsymbol{Z})$ to ensure unbiased item and person parameter estimation. If $\boldsymbol{Z}$ is excluded, then the likelihood of the observed item responses and the missing pattern $\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ is not proportional to a simple product of factorized densities.

$$\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi}) d\boldsymbol{Y}_{mis}. \tag{4.61}$$

The term $g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\phi})$ referring to the model of missingness cannot be brought out of the integral. ML estimation based on the observed item responses alone using $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ is potentially biased and ML inference might thus be invalid.

ML estimation in IRT models if the missing mechanism is MAR given $\boldsymbol{Z}$ As a special MAR condition, the missing data mechanism w.r.t. $\boldsymbol{Y}$ has been defined to be missing at random given $\boldsymbol{Z}$ if conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$ applies (see Equation 2.18). Conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Z}$ is implied by this definition although conditional stochastic dependence $\boldsymbol{D} \not\perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ applies. The full likelihood can then be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, z; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Z} = z; \boldsymbol{\phi}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Z} = z; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Z} = z; \boldsymbol{\phi}). \end{aligned} \tag{4.62}$$

As in the case of MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$, the observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ is sufficient for unbiased ML estimation with respect to the item and person parameters of the measurement model of $\boldsymbol{\xi}$. Here, as well, $\boldsymbol{Z}$ cannot be left out in the item and person parameter

estimation since Equation 4.61 applies ML estimation as a statistical inference based on $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ is not trustworthy.

ML estimation in IRT models if the missing mechanism is MAR given $\boldsymbol{Y}$ The third MAR condition is the conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$. In this case, the missing data mechanism w.r.t. $\boldsymbol{Y}$ is called missing at random given $\boldsymbol{Y}$ (see Equation 2.19). The full likelihood of the observed data is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, z; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}). \qquad (4.63)
\end{aligned}$$

Even if the covariate $\boldsymbol{Z}$ is left out, it follows

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\
&\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}; \boldsymbol{\phi}). \qquad (4.64)
\end{aligned}$$

Hence, both observed data likelihoods $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$ and $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ including the covariate are sufficient for valid direct likelihood inference and unbiased ML estimation with respect to item and person parameters of the measurement model of $\boldsymbol{\xi}$. Hence, the covariate $\boldsymbol{Z}$ can be included if this is of the researcher's substantial interest. However, is not required to fit a joint model of $(\boldsymbol{Y}, \boldsymbol{Z})$ in order to obtain unbiased estimates of persons' ability and item parameters.

ML estimation in IRT models if the missing mechanism is MNAR If the missing data mechanism w.r.t. $\boldsymbol{Y}$ is missing not at random as defined in section 2.2, then conditional stochastic dependence $\boldsymbol{D} \not\perp \boldsymbol{Y}_{mis} \,|\, (\boldsymbol{Y}_{obs}, \boldsymbol{Z})$ applies. Equation 4.57 cannot be further simplified. The conditional distribution $g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Z} = z, \boldsymbol{Y}_{mis}; \boldsymbol{\phi})$ cannot be placed outside the integral over $\Omega_{\boldsymbol{Y}_{mis}}$. Neither the use of the observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota}, \boldsymbol{\phi})$ nor $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota}, \boldsymbol{\phi})$ yield unbiased ML estimation. Statistical inference is therefore not trustworthy. Systematic biases of parameter estimates as demonstrated in Chapter 3 cannot be ruled out. If the missing data mechanism is MNAR, then the missing data mechanism is nonignorable and $\boldsymbol{D}$ needs to be included in the model and the full data

likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$ needs to be maximized.

4.5.2 IRT Models for Ignorable Missing Data

This work focuses on model-based methods for nonignorable missing data. Nevertheless, it is worthwhile to consider models for missing data that are MAR because of their structural resemblance to models for nonignorable item nonresponses, which are developed below. It was repeatedly stressed that the term *ignorable* is rather unfortunate. Strictly speaking, missing data are always problematic and should never be ignored. As previously mentioned, the term *noninformative* is synonymous with ignorable. If a missing data mechanism is noninformative, missingness does not contain additional information about unobserved quantities such as missing data and the parameters aimed to be estimated. For application, this implies that $\boldsymbol{D}$ needs not to be included in the parameter estimation of $\boldsymbol{\iota}$ to yield unbiased estimates. Hence, $\boldsymbol{D}$ can be ignored but not the missing data itself. Even if the missing data mechanism is MCAR or MAR, appropriate missing data methods are typically required to account for nonresponses. Even if the missing data mechanism is MCAR, approaches such as listwise deletion may be not applicable since no case with complete data exists. This is common in educational testings with planned missing data due to item sampling designs(Frey et al., 2009; Johnson, 1992; T. E. Raghunathan & Grizzle, 1995). With EM-algorithm (McLachlan & Krishnan, 2008, 2008) and FIML estimation (Arbuckle, 1996; Enders, 2001a), techniques exist that allow for unbiased parameter estimation if the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR. In many applications, researchers have to include covariates if the missing data mechanims w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$. If these covariates are not part of the substantial model, which is commonly the case, they are called *auxiliary* variables. A joint model of $\boldsymbol{Y}$ and auxiliary variables $\boldsymbol{Z}$ needs to be specified that preserve the target model of $\boldsymbol{Y}$. This modeling task can be challenging. In SEM the so-called *Spider model* proposed by Graham (2003) is a well-known and widely used model to include auxiliary variables in FIML estimation. Covariates and auxiliary variables can be left out if the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y}$. FIML will yield unbiased ML estimates based on the specified target model alone. FIML has become a state-of-the-art method in linear SEM (Schafer & Graham, 2002). The term FIML is rarely used in the context of IRT parameter estimation with missing data. However, standard JML and MML estimation methods as implemented in most common IRT software can be regarded as FIML estimation. Using two examples, this will be briefly explained here. First, computerized adaptive testing (CAT) is considered, and, second, two-stage testing with a routing test will be examined.

Computerized Adaptive Testing (CAT) In computerized adaptive testings, only a subset of items is answered by test takers until an a priori defined stop criterion is reached. If the starting item is fixed for all test takers or randomly chosen out of an item bank, then the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y}$ (Mislevy & Wu, 1996). After initial item responses, the following items are assigned depending on previous *observed* response behavior. Hence, $P(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}) = P(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{\boldsymbol{obs}})$. As Glas (2006) showed, IRT parameter estimation using common MML estimation is unbiased. Obviously, MML estimators as implemented in standard IRT software functions as FIML estimators. The reason is that MML estimation does not depend on bivariate summary statistics such as covariances between the items. Each observed item response is included in the ML function. If the CAT was applied to a sample of N test takers, then the observed data likelihood $\mathcal{L}(y_{obs}; \boldsymbol{\iota})$ ignoring the missing data mechanism (see Equation) can be written as

$$\begin{aligned} \mathcal{L}(y_{obs}; \boldsymbol{\iota}) &\propto \prod_{n=1}^{N} P(\boldsymbol{Y}_{\boldsymbol{n;obs}} = \boldsymbol{y}_{\boldsymbol{n;obs}}; \boldsymbol{\iota}) \\ &\propto \prod_{n=1}^{N} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{D_{ni}}. \end{aligned} \tag{4.65}$$

Hence, the response indicator variables D_i function as selecting variables that determine which elements of the complete data matrix $\boldsymbol{Y} = \boldsymbol{y}$ are observable. From Equation 4.65 follows that JML estimation yields also unbiased parameter estimates. In MML estimation, the unconditional probability of each observed response pattern is modeled by integration over the distribution of the latent ability $\boldsymbol{\xi}$. Hence, the MML function results from Equation 4.65 by

$$\mathcal{L}(\boldsymbol{y}_{\boldsymbol{obs}}; \boldsymbol{\iota}) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{D_{ni}} d\boldsymbol{\xi}.$$

It is important to note that the use of auxiliary information in CAT leads to a violation of ignorability conditions in CAT. As Glas (Glas, 2006) demonstrated, item and person parameter estimation is potentially biased if additional information such as educational achievements, socioeconomic status, etc. are used to determine the starting items of the CAT. In this case typically $\boldsymbol{D} \not\perp \boldsymbol{Y}_{\boldsymbol{mis}} \,|\, \boldsymbol{y}_{\boldsymbol{obs}}$. Missingness depends also on additional variables that are used to choose initial items. Formally, these variables can considered to be covariates $\boldsymbol{Z}$ in this CAT design, implying that $\boldsymbol{D} \perp \boldsymbol{Y}_{\boldsymbol{mis}} \,|\, (\boldsymbol{y}_{\boldsymbol{obs}}, \boldsymbol{Z})$. Hence, the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$. ML parameter estimation based on $\mathcal{L}(\boldsymbol{y}_{\boldsymbol{obs}}; \boldsymbol{\iota})$

is potentially biased. ML estimation needs to be based on $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ to ensure unbiasedness. The covariate $\boldsymbol{Z}$ is an auxiliary variable. The target model - here the measurement model of $\boldsymbol{\xi}$ - needs to be preserved in the joint model of $(\boldsymbol{Y}, \boldsymbol{Z})$. This problem is essentially equivalent to the use of a routing test to determine the final test form for test takers in paper and pencil tests. This example will be considered next.

Routing Tests The reliability of person parameter estimates using IRT models is not merely determined by the number of administered items, but also influenced by the item parameters. The test information $\boldsymbol{I}(\xi)$ is high and, therefore, the standard errors of $\hat{\xi}$ are small if the item difficulties and the persons' ability levels are close together (see Section 3.3). This is the rationale underlying CAT as well as branched testing and multistage testing (Lord, 1980). In contrast to CAT, the number of administered items is fixed in multistage testing. However, the particular selection of items presented to each test taker is not determined in advance. The simplest form is a two-stage testing design with a routing test administered firstly to obtain a rough impression of persons' ability levels and to determine the appropriate final test form used for parameter estimation. Formally, the routing test is an auxiliary variable. Together with other additional variables, the test score of the routing test constitutes the covariate $\boldsymbol{Z}$. The not-administered items in the final test $\boldsymbol{Y}$ are missing depending on $\boldsymbol{Z}$. The missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$, since $\boldsymbol{D} \perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$. From Equation 4.62 follows that ML estimation needs to be based on $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ instead of $\mathcal{L}(\boldsymbol{y}_{obs}; \boldsymbol{\iota})$. The crucial question is, how to included $\boldsymbol{Z}$ in a joint model of $(\boldsymbol{Y}, \boldsymbol{Z})$ preserving the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$. One approach is to include $\boldsymbol{Z}$ in a background model (DeMars, 2002). That is, a latent regression model $E(\boldsymbol{\xi} \,|\, \boldsymbol{Z})$ is included. Using MML estimation, the parameter vector $\boldsymbol{\iota} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Psi})$ contains additional quantities: (a) the parameter $\boldsymbol{\Gamma}$ of regression coefficients including the intercepts and (b) $\boldsymbol{\Psi}$ the variance-covariance matrix of latent residual $\boldsymbol{\zeta} = \boldsymbol{\xi} - E(\boldsymbol{\xi} \,|\, \boldsymbol{Z})$. However, the item parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ remain estimable parameters in this joint model. How can this model be justified theoretically? From Equation 4.62 merely follows that the likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ of the observed data is proportional to the joint distribution $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, z; \boldsymbol{\iota})$ that can be factorized in different ways. Assuming that the cases, and

therefore the rows of $\boldsymbol{Y}$, are independent, the following factorization can be considered:

$$\begin{aligned} \mathcal{L}(\boldsymbol{y_{n;obs}}, z; \boldsymbol{\iota}) &\propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} g(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}}, \boldsymbol{Z}_n = z_n, \boldsymbol{\xi}; \boldsymbol{\iota}) d\boldsymbol{\xi} \\ &\propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} g(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \boldsymbol{Z}_n = z_n, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{Z}_n = z_n, \boldsymbol{\xi}; \boldsymbol{\iota}) d\boldsymbol{\xi}. \end{aligned} \tag{4.66}$$

If no DIF exists with respect to $\boldsymbol{Z}$, that is

$$\begin{aligned} \mathcal{L}(\boldsymbol{y_{obs}}, z; \boldsymbol{\iota}) &\propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} g(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \,|\, \boldsymbol{Z}_n = z_n; \boldsymbol{\iota}) g(\boldsymbol{Z}_n = z_n) d\boldsymbol{\xi} \\ &\propto \prod_{n=1}^{N} g(\boldsymbol{Z}_n = z_n) \int_{\mathbb{R}^m} g(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \,|\, \boldsymbol{Z}_n = z_n; \boldsymbol{\iota}) d\boldsymbol{\xi} \end{aligned} \tag{4.67}$$

The covariate $\boldsymbol{Z}$ is purely exogenous in this model. As in any regression model, the unconditional distribution of the predictor variables, here $g(\boldsymbol{Z}_n = z_n)$, can be left out of the ML estimation equation. Essential is the model of the conditional distribution of the endogenous variables given the independent variables. Assuming local stochastic independence for all manifest variables Y_i, the MML estimation equation can be written as

$$\begin{aligned} L(\boldsymbol{y_{obs}}, z; \boldsymbol{\iota}) &\propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} g(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \,|\, \boldsymbol{Z}_n = z_n; \boldsymbol{\iota}) d\boldsymbol{\xi} \\ &\propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{D_{ni}} g(\boldsymbol{\xi} \,|\, \boldsymbol{Z}_n = z_n; \boldsymbol{\iota}) d\boldsymbol{\xi}. \end{aligned} \tag{4.68}$$

Hence, given no DIF exists, auxiliary variables can be included as independent variables in a latent regression model. The regression parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ determines the conditional distribution $g(\boldsymbol{\xi} \,|\, \boldsymbol{Z} = z; \boldsymbol{\iota})$. In application, it is commonly assumed that the latent residual is multivariate normal with $\boldsymbol{\zeta} \sim (\boldsymbol{0}, \boldsymbol{\Psi})$. Software like ConQuest(Wu et al., 1998) or M*plus* (Muthén & Muthén, 1998 - 2010) allow to estimate latent regression models and measurement models contemporarily using Marginal ML estimation. Figure 4.13 illustrates the model with a unidimensional latent variable ξ and a single covariate Z in a latent regression model. It can be seen that measurement invariance with respect to Z is assumed, so that $Y_i \perp Z \,|\, \boldsymbol{\xi}$ for al Y_i.

In Section 2.3, it was shown that $D_i \perp Y_i \,|\, \boldsymbol{Z}$ implies $D_i \perp \boldsymbol{\xi} \,|\, \boldsymbol{Z}$ given the missing data mechanism with respect to Y_i is MAR given $\boldsymbol{Z}$ and measurement invariance hold. Con-

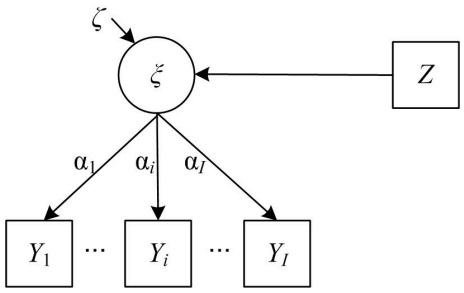

Figure 4.13: Path diagram of a latent regression model for item nonresponses that are MAR given $\boldsymbol{Z}$.

sidering the whole test instead of single items, that is $\boldsymbol{D} \perp \boldsymbol{\xi} \,|\, \boldsymbol{Z}$. If item nonresponses result only from not-administered items due to the routing test, then this is trivial since $\boldsymbol{D}$ is a function $f(\boldsymbol{Z})$ of the auxiliary variables. Each value $\boldsymbol{z}$ is associated with a certain missing pattern $\boldsymbol{d}$. Based on these considerations it can be explained how the latent regression model accounts for missing data. If $\boldsymbol{Z}$ includes a very reliable and valid routing test, then $\boldsymbol{\xi}$ and $\boldsymbol{Z}$ will be strongly stochastically dependent. The stronger the stochastic dependence is, the better the routing test works. This implies that $\boldsymbol{\xi}$ and $\boldsymbol{D}$ are also strongly stochastically dependent on each other. The ability distributions will considerably differ between the missing pattern $\boldsymbol{D} = \boldsymbol{d}$. If this information about the systematic differences in the latent ability distribution dependending on the missing pattern is ignored, then ML estimates of IRT parameters are potentially biased. The inclusion of a latent regression model accounts for different distributions of $\boldsymbol{\xi}$ given $\boldsymbol{D}$ by allowing for distributional differences of $g(\boldsymbol{\xi} \,|\, \boldsymbol{Z} = z; \boldsymbol{\iota})$ for different values $\boldsymbol{Z} = z$. More specifically, the values $E(\boldsymbol{\xi} \,|\, \boldsymbol{Z} = z)$ of the latent regression vary depending on $\boldsymbol{Z}$, allowing for different average ability levels. However, only one variance-covariance matrix $\boldsymbol{\Psi}$ is estimated. Thus, equal variance-covariance structures are assumed for all values $\boldsymbol{Z} = z$. Multiple group models might be a less restrictive alternative when $\boldsymbol{Z}$ is discrete.

However, CAT and missing responses due to item selection based on routing tests were discussed exemplarily. Many other cases could be considered with ignorable missing data. For example, Glas (1988) showed that MML estimation is unbiased in multistage testings if the items of the routing test are indicators of exactly the same latent ability $\boldsymbol{\xi}$ as indicated by $\boldsymbol{Y}$. In this case, the measurement model of $\boldsymbol{\xi}$ can be based on $(\boldsymbol{Y}, \boldsymbol{Z})$, dispensing with the need of the latent regression model. Unfortunately, missing responses are very likely nonignorable in many applications. Omitted and not-reached items, for example, occur more likely in persons with lower proficieny levels (Culbertson, 2011, April; Rose et al., 2010). Hence, missingness depends most likely on unobserved variables such as the latent

ability $\boldsymbol{\xi}$. In such cases, the missing data mechanism w.r.t. $\boldsymbol{Y}$ is NMAR. In the following sections, appropriate IRT model based approaches for non-ignorable missing data will be examined and further developed.

4.5.3 Multidimensional IRT Models for Non-ignorable Missing Data

Given the missing data mechanism is MAR, a lot of approaches have been developed in order to obtain unbiased estimates and correct statistics. In contrast, if the missing data mechanism is non-ignorable, then only few approaches exist. It has turned out to be challenging to find a general approach for non-ignorable missing data. Up to now, two general classes of models have been proposed: (a) Selection Models (SLM) (Heckman, 1976, 1979; Amemiya, 1984; Little, 1993, 1995; Little & Rubin, 2002; Enders, 2010) and (b) Pattern Mixture Models (PMM) (Glynn et al., 1986; Little, 1993, 1995; Little & Rubin, 2002; Little, 2008). Since missingness is informative with respect to unobserved variables $\boldsymbol{Y}_{mis}$ and, therefore, to the unknown parameters $\boldsymbol{\iota}$ underlying $\boldsymbol{Y}$, the missing indicator variable $\boldsymbol{D}$ needs to be included in a joint model of $(\boldsymbol{Y}, \boldsymbol{D})$. This is the underlying rationale of both SLM and PMM. Hence, ML estimation in both classes of models is based on the joint distribution $g(\boldsymbol{Y}, \boldsymbol{D})$ of these variables given a particular model.

In the recent years, MIRT models for nonignorable missing data have been proposed by O'Muircheartaigh and Moustaki (1999), Moustaki and Knott (2000), Holman and Glas (2005), Korobko et al. (2008), Glas and Pimentel (2008), and Rose, von Davier & Xu (2010). These models can be derived from both SLM and PMM under particular assumptions. In this chapter, MIRT models for missing responses are developed from the general SLM. Heckman's SLM (Heckman, 1976, 1979) for normally distributed variables Y is used to introduce SLM in general. Based on these considerations, appropriate IRT models for item nonresponses will be derived step by step.

4.5.3.1 MIRT Models as Likelihood Based Missing Data Method

In Section 4.5.1, ML estimation with missing data was scrutinized. It could be shown that $\boldsymbol{D}$ can be ignored in ML estimation procedures if two ignorability conditions hold: (a) the nonresponse mechanism w.r.t. $\boldsymbol{Y}$ is MCAR or MAR and (b) distinctness of the parameter spaces $\Omega_{\boldsymbol{\iota}}$ and $\Omega_{\boldsymbol{\phi}}$[12]. In many applications, the ignorability assumptions are unlikely to hold true. Classical examples are described in clinical trials, where attrition

[12]Cases can be constructed where the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR but distinctness of $\Omega_{\boldsymbol{\iota}}$ and $\Omega_{\boldsymbol{\phi}}$ does not hold. These cases are not considered here.

is often caused by severe aggravation or even death in study participants (e. g. Enders, 2010; Pauler, McCoy, & Moinpour, 2003). In educational assessments, strong empirical evidence was repeatedly found that unplanned missing are associated with the persons' proficiency levels (Rose et al., 2010; Culbertson, 2011, April; McCaffrey & Lockwood, 2011). Suited approaches to handle missing responses are required.

In educational and psychological large scale assessments, IRT models are commonly used to obtain item parameters and to quantify persons' proficiency levels. Maximum likelihood estimation is the most popular method used for parameter estimation (Baker & Kim, 2004). In Section 4.5.1, ML estimation with missing data was examined. The Equations 4.49 - 4.51 illustrate the difference between ML estimation given the missing data mechanism is MAR or NMAR. In the latter case, the likelihood function cannot be factorized into two independent pieces, which refer to the target model of $\boldsymbol{Y}$ and the missing data model of $\boldsymbol{D}$. As a consequence, unbiased ML estimation of parameters $\boldsymbol{\iota}$ in presence of nonignorable missing data needs to be based on a joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$. The ML functions are then proportional proportional to the joint distribution $g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$. This is the underlying rationale of SLMs as well as PMMs that will be introduced next.

SLM versus PMM In general, the ML function is proportional to the joint distribution that can be factorized in different ways. In contrast to classical literature, here it is distinguished between $\boldsymbol{Y}$ the items that constitute the measurement model of $\boldsymbol{\xi}$ and $\boldsymbol{Z}$ the multidimensional covariate. If $\boldsymbol{Z}$ is a purely exogenous variable, then the joint distribution $g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Z} = z; \boldsymbol{\iota}, \boldsymbol{\phi})$ needs to be considered. The complete data likelihood with respect to $(\boldsymbol{\iota}, \boldsymbol{\phi})$ is then proportional to

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}, z; \boldsymbol{\phi}, \boldsymbol{\iota}) \propto g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = z; \boldsymbol{\phi}) g(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{Z} = z; \boldsymbol{\iota}), \tag{4.69}$$

or alternatively

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}, z; \boldsymbol{\phi}, \boldsymbol{\iota}) = g(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{Z} = z; \boldsymbol{\iota}) \cdot g(D = d \,|\, \boldsymbol{Z} = z; \boldsymbol{\phi}). \tag{4.70}$$

The two likelihood functions refer to both major modelling approaches for nonignorable missing data - SLMs are represented by Equation 4.69 and PMMs by Equation 4.70. It can be seen that the two parameter vectors $\boldsymbol{\phi}$ and $\boldsymbol{\iota}$ refer to the two parts of the joint model. $\boldsymbol{\iota}$ is of substantive interest, whereas $\boldsymbol{\phi}$ is a nuisance but required to be jointly estimated with $\boldsymbol{\iota}$. In applications, the conditional distributions are model by appropriately.

Assuming there is only a unidimensional variable Y with its response indicator D, the conditional distribution $g(D = d \mid Y = Y, \boldsymbol{Z} = z; \boldsymbol{\phi})$ in SLM can be modeled by a probit or logit regression $P(D = 1 \mid Y = y, \boldsymbol{Z} = z; \boldsymbol{\phi})$. In this case, the parameter vector $\boldsymbol{\phi}$ consists of the regression coefficients and the threshold or the intercept term of the this regression model. In PMMs, $\boldsymbol{\phi}$ is of minor interest. The estimate $\hat{\boldsymbol{\iota}}$ is obtained by averaging over all pattern-specific estimates $\hat{\boldsymbol{\iota}}_{\boldsymbol{d}}$. That is, in a first step the model parameters $\hat{\boldsymbol{\iota}}$ are estimated within each observed missing pattern $D = d$. Subsequently, the estimates $\hat{\boldsymbol{\iota}}_d$ are combined to a single estimate $\hat{\boldsymbol{\iota}}$ (Enders, 2010). Little (Little, 2008) noted that PMMs are notoriously underidentified. The estimates $\hat{\boldsymbol{\iota}}_{\boldsymbol{d}}$ are not estimable without restrictions resting upon typically untestable assumptions (2001; 2010). However, identification issues are also challenging in many applications of selection models. In the classical normal SLM proposed by Heckman, the model is only identified by very strong distributional assumptions. A comprehensive comparison of SLMs and PMMs is beyond the scope of this work. Little (Little, 2008) argued that the factorization underlying the joint distribution that underlies SLMs are a more natural choice. For that reason SLMs are focused on here.

Derivation of the MIRT model for nonignorable missing data In this section, a multidimensional IRT model for nonignorable missing data will be derived, which can be regarded as a selection model. The classical SLM introduced by Heckman (Heckman, 1976, 1979) rests upon very strong normality assumptions. Furthermore, it is appropriate in regressions with only a single manifest dependent variable Y. In IRT measurement models, there is a multivariate dependent variable $\boldsymbol{Y} = Y_1, \ldots, Y_I$. Accordingly, the response indicator variable $\boldsymbol{D}$ is multidimensional as well. Each item Y_i is non-normally distributed. In the case of dichotomous items, each Y_i unconditionally and conditionally Bernoulli distributed. Hence, some generalizations and modifications of classical SLMs are required to be appropriate for IRT measurement models. The parameter vector $\boldsymbol{\iota}$ consists of the item and person parameters in the case of JML, and of the item parameters and the parameters describing the distribution of $\boldsymbol{\xi}$ in the case of MML. The inclusion of a model for $\boldsymbol{D}$ should correct for nonignorable missing data without changing the substantial model of interest given by the parameter vector $\boldsymbol{\iota}$. A fundamental difference to commonly used selection models is that the independent variable of the regressions $P(Y_i = 1 \mid \boldsymbol{\xi})$ is a latent variable, which is always missing. As previously noted, $\boldsymbol{\xi}$ is constructed in a measurement model rather than simply measured. The extension of the model by including a model for $\boldsymbol{D}$ needs to preserve the measurement model of $\boldsymbol{\xi}$ in the sense that the construction of $\boldsymbol{\xi}$ remains unaffected. Several generalizations of SLMs have already been introduced in the

past. Dubin and Rivers (1989) generalized the Tobit model and Heckman's SLM to the case of discrete manifest variables Y (Dubin & Rivers, 1989). Generally, extensions of SLM using ML estimation are natural for dependent variables with distributions of their residuals that are among the exponential family (Barndorff-Nielsen, 1976). Therefore, SLM can be extended to many models which belong to the class of generalized linear models.

MIRT models for nonignorable missing data are a consequential further development of SLMs for item nonresponses in measurement models with categorical manifest variables. As mentioned above,in all SLMs and PMMs the model identification is a problematic issue. PMMs are never identified without untestable assumptions. SLMs are weakly identified (e. g. Little, 2008) and tend to suffer from convergence problems (e. g. Toomet & Henningsen, 2008). The basic idea underlying MIRT models for nonignorable missing data is the construction of a latent response propensity θ underlying the response indicator variables $D_1, \ldots, D_I$ (O'Muircheartaigh & Moustaki, 1999; Holman & Glas, 2005; Moustaki & Knott, 2000; Korobko et al., 2008; Glas & Pimentel, 2008). Like $\boldsymbol{\xi}$, θ is defined as a function $f(U)$ of the person variable U. Hence, the MIRT model for nonignorable missing responses is a joint measurement model of the latent ability $\boldsymbol{\xi}$ and the latent response propensity θ based on $(\boldsymbol{Y}, \boldsymbol{D})$. The major advantage of this model is that it does not suffer from identification problems. The model equations of the measurement model of $\boldsymbol{\xi}$ are given by the parametric regressions $P(Y_i \,|\, \boldsymbol{\xi})$ (e. g. Rasch model, Birnbaum model, or 3PL-model). Similarly, there are I regressions $P(D_i \,|\, \theta)$ constituting the measurement model of the latent response propensity. Typically, the 1PLM or the 2PLM is chosen to model $P(D_i \,|\, \theta)$.

The measurement model of θ can be utilized in two different ways : First, in order to estimate item response propensities $P(Y_i \,|\, \theta)$ for each examinee and each item that serve to construct weights, which can be used in a subsequent estimation of the measurement model of $\boldsymbol{\xi}$. This approach belongs to weighting procedure. The problem is that each observed item response needs to be weighted differently, even for a single test taker. However, to the best knowledge of the author no IRT software exists that allows for more than one weight per observational unit. Therefore, the application of weighting procedures for item nonresponses is not possible yet. Nevertheless, using weights for each single response is possible at least theoretically and may be worth considering in future research.

A second approach is to include the measurement model of θ in a joint multidimensional IRT model with θ and $\boldsymbol{\xi}$. The data matrix used for estimation of this MIRT model is the combined $N \times 2I$ matrix $(\boldsymbol{Y}, \boldsymbol{D}) = (\boldsymbol{y}, \boldsymbol{d})$ of the items and the respective response in-

dicators. The weighting approach is a two-step procedure with the item response propensities $P(Y_i = 1 \mid \theta)$ taken as fixed in the estimation of the measurement model of $\boldsymbol{\xi}$. In contrast, the MIRT approach requires a more complex model but all parameters can be estimated simultaneously in one step. Additionally, as by-product parameters, some of the parameters in the MIRT model allow to quantify the relationship between the latent variable $\boldsymbol{\xi}$ and the probability of item non-responses. Thus, this model can provide additional diagnostic value as well. The univariate normal SLM was based on the factorization $g(Y, D \mid Z; \boldsymbol{\iota}, \boldsymbol{\phi}) = g(Y \mid Z; \boldsymbol{\iota}) g(D \mid Y, Z; \boldsymbol{\phi})$. The MIRT model for nonignorable missing data as proposed by O'Muircheartaigh and Moustaki (1999), Moustaki and Knott (2000), Holman and Glas (2005), Korobko et al. (2008), and Glas and Pimentel (2008) did not involve additional covariates $\boldsymbol{Z}$. Rose, von Davier and Xu (2010) included categorical covariates using multiple group MIRT model for nonignorable missing data. For the beginning, the covariate $\boldsymbol{Z}$ will be left out in order to introduce the basic MIRT model for missing data that are NMAR. In this case, the model can be derived from the basic factorization $g(\boldsymbol{Y}, \boldsymbol{D}) = g(\boldsymbol{Y}) g(\boldsymbol{D} \mid \boldsymbol{Y})$. Similarly to univariate SLM, a parametric model is chosen for the joint distribution so that $g(\boldsymbol{Y}, \boldsymbol{D}; \boldsymbol{\iota}, \boldsymbol{\phi}) = g(\boldsymbol{Y} \boldsymbol{\iota}) g(\boldsymbol{D} \mid \boldsymbol{Y}; \boldsymbol{\phi})$. Using the partition $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$, the observed data likelihood for a sample size of N with independent cases n is

$$\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} g(\boldsymbol{Y}_{obs;n} = \boldsymbol{y}_{obs;n}; \boldsymbol{\iota}) g(\boldsymbol{D}_n = \boldsymbol{d}_n \mid \boldsymbol{Y}_{obs;n} = \boldsymbol{y}_{obs;n}; \boldsymbol{\phi}) \quad (4.71)$$

The likelihood consists of two parts referring to the two models indexed by $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$. Hence, Equation 4.71 seems to be equal to Equation 4.51. However, the relation between the observed data likelihood and the theoretical complete data likelihood are different for the different missing data mechanisms (cf. 4.51 and 4.50). The independent maximization of the likelihood with respect to $\boldsymbol{\iota}$ omitting the model of $\boldsymbol{D}$ would result in biased ML estimates given the missing data mechanism is nonignorable (see Section 4.5.1). Formally, the latent variable $\boldsymbol{\xi}$ can be regarded as an estimable parameter of the vector $\boldsymbol{\iota}$. Similarly, the latent response propensity can be considered to be part of the parameter vector $\boldsymbol{\phi}$ [13]. In contrast, $\boldsymbol{\xi}$ and θ are treated as random variables in commonly used MML estimation procedures. In the further derivations the person variables $\boldsymbol{\xi}$ and θ and parameter vectors $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$ are written separately to keep in line with commonly used notation for latent trait

[13] Considering individual values of the latent variables of each test taker as fixed and estimable model parameter refers to the fixed effects approach that underlies JML and CML estimation.

models. Hence, Equation 4.71 can be written as

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} g(\boldsymbol{Y_{obs;n}} = \boldsymbol{y_{obs;n}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{D}_n = \boldsymbol{d}_n \,|\, \boldsymbol{Y_{obs;n}} = \boldsymbol{y_{obs;n}}, \boldsymbol{\xi}, \boldsymbol{\theta}; \boldsymbol{\phi}) \quad (4.72)$$

In contrast to most papers using MIRT models for missing data, here the latent response propensity can be a p-dimensional latent variable $\boldsymbol{\theta} = \theta_1, \ldots, \theta_p$ with $\Omega_{\boldsymbol{\theta}} = \mathbb{R}^p$. In most applications, θ is assumed to be unidimensional. In Section 4.5.3.4, it will be shown that the correct dimensionality of the latent response propensity is of major importance and needs to be carefully examined. Up to now, this issue has not been sufficiently addressed in the literature. Existing MIRT models for nonigorable missing data rest upon the assumption of local stochastic independence , similar to those of common IRT models. In particular, these are:

$$Y_i \perp (\boldsymbol{Y_{-i}}, \boldsymbol{D}, \boldsymbol{\theta}) \quad | \quad \boldsymbol{\xi} \qquad \forall i = 1, \ldots, I \quad (4.73)$$

$$D_i \perp (\boldsymbol{D_{-i}}, \boldsymbol{Y}) \quad | \quad (\boldsymbol{\xi}, \boldsymbol{\theta}) \qquad \forall i = 1, \ldots, I, \quad (4.74)$$

The first assumption implies $P(Y_i = y_i \,|\, \boldsymbol{\xi}, \theta; \boldsymbol{\iota}) = P(Y_i = y_i \,|\, \boldsymbol{\xi}, \boldsymbol{\iota})$. This assumption of local stochastic independence is crucial, since it refers to the construction of $\boldsymbol{\xi}$. Note that the measurement model comprises not only test items Y_i, but also response indicators D_i. The inclusion of the latter should correct for missingness in the estimation of model parameters $\boldsymbol{\iota}$ but should not change the substantive meaning of item and person parameters. Assumption 4.5.4 is necessary to ensure that $\boldsymbol{\xi}$ in the MIRT models for nonignorable missing data is constructed in same way as in the unidimensional IRT model if no missing data exist. In conjunction with assumption 4.74, it follows that all manifest variables D_i and Y_i are conditionally stochastically independent from each other given the latent variables in the MIRT model. Based on these assumptions, Equation 4.72 can be written as

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \quad \propto \quad \prod_{n=1}^{N} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} P(D_{ni} = d_{ni} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}; \boldsymbol{\phi}). \quad (4.75)$$

Since $\boldsymbol{\xi}$ is a m-dimensional variable defined in $\Omega_{\boldsymbol{\xi}} = \mathbb{R}^m$ and $\boldsymbol{\theta}$ is a P-dimensional variable in $\Omega_{\boldsymbol{\theta}} = \mathbb{R}^p$, the MML function of the observed data is given by

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \quad \propto \quad \prod_{n=1}^{N} \int_{\mathbb{R}^m} \int_{\mathbb{R}^p} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} P(D_{ni} = d_{ni} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}; \boldsymbol{\phi}) g(\boldsymbol{\xi}, \boldsymbol{\theta}), \quad (4.76)$$

where the joint distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ of the latent variables is typically chosen to be a multivariate normal. Note that the exponent d_{ni} selects the observed variables Y_{ni} that are part of the observed data likelihood [14]. Hence, only the observed item responses y_{ni}, indicated by $d_{ni} = 1$, can be included in parameter estimation of the measurement model of $\boldsymbol{\xi}$. The likelihood functions 4.75 and 4.76 represent the general MIRT model for the nonignorable model that was derived from the general SLM by the constrction of a latent response propensity and certain assumptions given by Equations 4.5.4 and 4.74. As O'Muircheartaigh and Moustaki (1999) demonstrated, the same MIRT model can alternatively be derived from the general PMMs based on the same assumptions.

Between- and within-item MIRT models for nonignorable missing data In the literature, different MIRT models for nonignorable item nonresponses have been developed that can be broadly divided into between-item multidimensional IRT (BMIRT) models and the within-item multidimensional IRT (WMIRT) models (Adams, Wilson, & Wang, 1997; Hartig & Höhler, 2008; Wang, Wilson, & Adams, 1997). Which of these models should be used in real applications? Here it is argued that both, BMIRT and WMIRT models, account equally well for nonignorable missing data, but the interpretation of some parameters differs between the two classes of models. Furthermore, it will be shown that WMIRT models are not necessarily equivalent to BMIRT models. The issue of model equivalence will be addressed in detail below. The general model equations of the manifest variables Y_i and D_i in all MIRT models discussed in this work will be introduced first. The multidimensional extension of the 2PLM for dichotomous items Y_i is chosen as the measurement model of both $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, which includes the multidimensional 1PLM as a special case (e. g. Embretson & Reise, 2000; Reckase, 1997). The model equation of the items Y_i is given by

$$P(Y_i = 1 \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) \;=\; \frac{\exp(\boldsymbol{\alpha_i}^T \boldsymbol{\xi} - \beta_i)}{1 + \exp(\boldsymbol{\alpha_i}^T \boldsymbol{\xi} - \beta_i)}. \tag{4.77}$$

If $\boldsymbol{\xi}$ is a M-dimensional latent variable, then $\boldsymbol{\alpha}_i$ is a vector with M item discriminations[15] $\alpha_{i1}, \ldots, \alpha_{im}, \ldots, \alpha_{iM}$. The model equation of the respective response indicators D_i is

$$P(D_i = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{\xi}; \boldsymbol{\phi}) \;=\; \frac{\exp(\boldsymbol{\gamma_i}(\boldsymbol{\xi}, \boldsymbol{\theta})^T - \gamma_{i0})}{1 + \exp(\boldsymbol{\gamma_i}(\boldsymbol{\xi}, \boldsymbol{\theta})^T - \gamma_{i0})}, \tag{4.78}$$

[14] The model equations with respect to missing responses $P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^0 = 1$ in the ML function do not affect the observed data likelihood.

[15] In within-item MIRT models the item discriminations are actually partial logistic regression coefficients. However, the term discrimination is conveniently retained here.

with $\boldsymbol{\gamma_i} = \gamma_{i1}, \ldots, \gamma_{im}, \ldots, \gamma_{iM}, \gamma_{i(M+1)}, \ldots, \gamma_{il}, \ldots, \gamma_{i(M+P)}$ as the vector of discrimination parameters and the thresholds γ_{i0}. If the 1PLM is used, then the elements in $\boldsymbol{\alpha_i}$ and $\boldsymbol{\gamma_i}$ can only take on the values zero and one. The choice between 1PLM or 2PLM needs to be answered individually in a particular application, depending on theoretical considerations, model fit, and potentially many other factors. For a clear distinction between the BMIRT and different WMIRT models, a general model equation in matrix notation is introduced. Let $\boldsymbol{l(Y, D)} = (l(Y_1), \ldots, l(Y_I), l(D_1), \ldots, l(D_I))$ be the vector of the logits in the MIRT model and $(\boldsymbol{\xi, \theta}) = (\xi_1, \ldots, \xi_P, \theta_1, \ldots, \theta_M)$ be the vector of latent variables. $\boldsymbol{\Lambda}$ is the $2I \times (P + M)$ matrix of discrimination parameters, and $(\boldsymbol{\beta, \gamma_0})$ is the vector of item difficulties and thresholds respectively. The multivariate logit model equation can be written as

$$\boldsymbol{l(Y, D)} \quad = \quad \boldsymbol{\Lambda}(\boldsymbol{\xi, \theta})^T - (\boldsymbol{\beta, \gamma_0})^T. \tag{4.79}$$

Rewriting this Equation reveals that matrix $\boldsymbol{\Lambda}$ consists of four blocks:

$$\begin{pmatrix} \boldsymbol{l(Y)} \\ \boldsymbol{l(D)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \boldsymbol{\gamma_\xi} & \boldsymbol{\gamma_\theta} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\theta} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma_0} \end{pmatrix}. \tag{4.80}$$

$\boldsymbol{l(Y)} = (l(Y_1), \ldots, l(Y_I))^T$ and $\boldsymbol{l(D)} = (l(D_1), \ldots, l(D_I))^T$ are the vectors of the respective logits. $\boldsymbol{\beta} = \beta_1, \ldots, \beta_I$ and $\boldsymbol{\gamma_0} = \gamma_{i0}, \ldots, \gamma_{I0}$ are the vectors of the item difficulties or threshold parameters of the variables Y_i and D_i. The matrix $\boldsymbol{\Lambda}$ consists of (a) $\boldsymbol{\alpha}$ the $I \times M$ matrix with the item discriminations α_{im}; (b) the $I \times M$ matrix $\boldsymbol{\gamma_\xi}$ consisting of the elements γ_{im} which relates the components ξ_m to the response indicators D_i; (c) the $I \times P$ matrix $\boldsymbol{\gamma_\theta}$ with the discrimination parameters γ_{il} that relate the latent dimensions θ_l and the response indicators D_i. In all MIRT models examined in this work, the upper right block in $\boldsymbol{\Lambda}$ needs to be a $I \times P$ zero matrix. This is essential to ensure that $\boldsymbol{\xi}$ is constructed equivalently in all MIRT models. Only in this case the meaning of $\boldsymbol{\xi}$ remain unchanged and the individual values of $\boldsymbol{\xi}$ as well as item parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) are comparable across alternative models. This is important since the measurement model of $\boldsymbol{\xi}$ is the target model which needs to be preserved as a part in a joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$ that accounts for missing data. Note that the vector $\boldsymbol{\gamma_i}$ of discrimination parameters of item i in Equation 4.78 is the i-th row of the submatrix $\boldsymbol{\gamma} = (\boldsymbol{\gamma_\xi}, \boldsymbol{\gamma_\theta})$, which is simply the $(M + i)$-th row of $\boldsymbol{\Lambda}$. In the following sections, the different MIRT models will be derived step by step starting with the BMIRT model for nonignorable missing responses. Afterwards, three equivalent WMIRT models will be developed rationally.

4.5.3.2 Between-item Multidimensional IRT Model for Nonignorable Missing Data

The terms between-item and within-item multidimensionality were introduced by Adams, Wilson, and Wang (1997) and Wang, Wilson, and Adams (1997). Between-item dimensionality is equivalent to simple structure in factor analytical terms. That is, each manifest variable indicates only one single latent dimension. Within-item dimensionality allows the items to be indicators of more than one latent dimension. Here in this work, the terms between- and within-item dimensionality are also used but in a less restrictive way. In the BMIRT model for nonignorable missing data, the assumption of conditional stochastic independence given by Equation 4.74 is modified, so that

$$D_i \perp (\boldsymbol{D}_{-i}, \boldsymbol{Y}, \boldsymbol{\xi}) \quad | \quad \boldsymbol{\theta} \qquad \forall i = 1, \ldots, I. \tag{4.81}$$

The second local stochastic assumption given by Equation 4.5.4 remains valid. From both assumptions follows that $\boldsymbol{D} \perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}$ and $\boldsymbol{Y} \perp \boldsymbol{\theta} \,|\, \boldsymbol{\xi}$, implying that matrix $\boldsymbol{\Lambda}$ of item discriminations in Equation 4.80 is block diagonal, so that

$$\begin{pmatrix} \boldsymbol{l}(\boldsymbol{Y}) \\ \boldsymbol{l}(\boldsymbol{D}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\gamma}_{\boldsymbol{\theta}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\theta} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma}_{\mathbf{0}} \end{pmatrix}. \tag{4.82}$$

Hence, $\boldsymbol{\gamma}_{\boldsymbol{\xi}} = \mathbf{0}$. That is why this model is labeled between-item multidimensional. In factor analytic terms; there are no cross factor loadings between $\boldsymbol{\xi}$ and the response indicators D_i due to conditional stochastic independence $D_i \perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}$. The model equation of the response indicators given by Equation 4.78 can be simplified to

$$P(D_i = 1 \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}) \quad = \quad \frac{\exp(\boldsymbol{\gamma}_{\boldsymbol{i};\boldsymbol{\theta}} \boldsymbol{\theta} - \gamma_{i0})}{1 + \exp(\boldsymbol{\gamma}_{\boldsymbol{i};\boldsymbol{\theta}} \boldsymbol{\theta} - \gamma_{i0})}. \tag{4.83}$$

It should be noted that *within* the measurement model of $\boldsymbol{\xi}$ the items Y_i can indicate more than one latent dimension ξ_m. Similarly, the response indicators D_i can indicate more than one latent dimension θ_l but none of the latent variables ξ_m. BMIRT models with such a complex dimensionality will be discussed below in detail (see page 205). Figure 4.14 displays a fictional example of a BMIRT model with a simplex dimensionality. The latent ability $\boldsymbol{\xi} = (\xi_1, \xi_2)$ and the latent response propensity $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are two-dimensional. Each item Y_i indicates only one latent dimension ξ_k, and each D_i indicates only one dimension θ_l. This implies a strong simple structure in the terminology of factor analysis (Thurstone, 1947). Note that the measurement model of $\boldsymbol{\theta}$ needs not to mimic the measurement model of $\boldsymbol{\xi}$. Hence, the parameter matrices of the item discriminations α_{ik} and

γ_{il} are not required to have the same structure or dimensionality. Even the number m of dimensions of $\boldsymbol{\xi}$ and the number p of dimensions of $\boldsymbol{\theta}$ are not required to be equal. The number of latent response propensities underlying $\boldsymbol{D}$ may depend on several factors and is not determined by the number of latent dimensions ξ_m as item positions, item types, and so on. Here it is argued that the dimensionality of $\boldsymbol{\theta}$ needs to be studied carefully. We will return to this point in Section 4.5.3.4. The advantage the BMIRT models is the easy

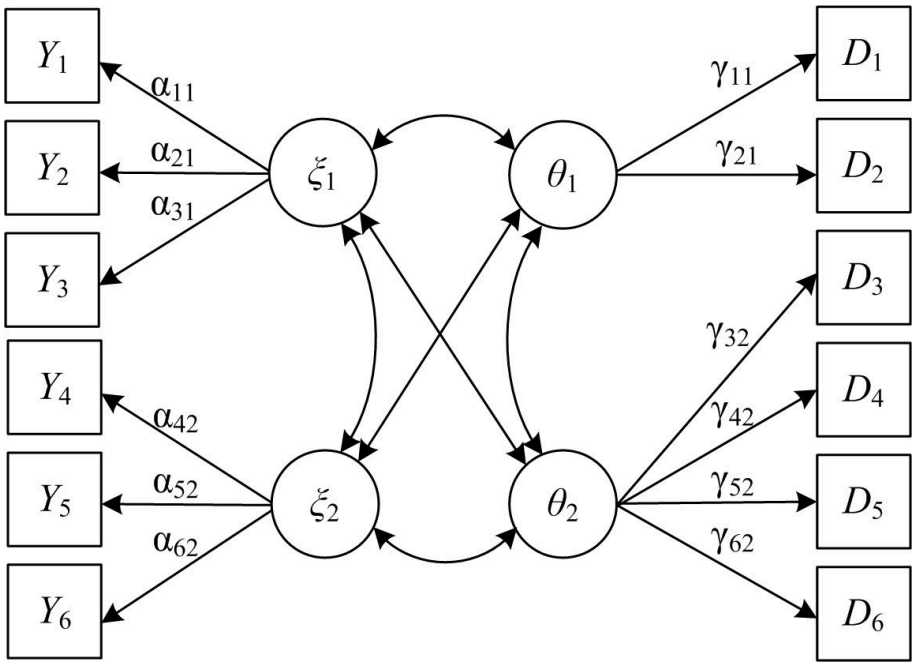

Figure 4.14: Graphical representation of the BMIRT model.

interpretation of the latent variables and item parameters. All dimensions ξ_k are scaled logits of the items Y_i that indicate ξ_k. Similarly, all latent variables θ_l are scaled logits of the respective response indicators D_i, which indicate θ_l. Therefore, all θ_l can indeed be interpreted as a latent response propensity in the sense that higher values of θ_l indicate a higher tendency to response to items D_i given $\gamma_{il} > 0$. The dimensions ξ_k are constructed in the same way as in a model without missingness. The meaning of these variables is unaffected. Higher values of ξ_k indicate higher probabilities to provide correct answers to test items Y_i if $\alpha_{ik} > 0$. The ease of the interpretation of the latent variables facilitates also the interpretation of the relationships between the latent dimensions. In commonly used MIRT models estimated by MML estimation, the joint distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ of the latent variables is assumed to be multivariate normal with the expected value $E(\boldsymbol{\xi}, \boldsymbol{\theta})$ and the variance-covariance matrix $\Sigma_{\boldsymbol{\xi},\boldsymbol{\theta}}$. In conjunction with the conditional stochastic in-

dependencies of the manifest variables given by Equations 4.5.4 and 4.81, covariances $Cov(\xi_k, \theta_l) \neq 0$ imply unconditional stochastic dependence $\boldsymbol{Y} \not\perp \boldsymbol{D}$. In turn, if the stochastic dependencies between all latent dimensions ξ_k and θ_l are linear and the $Cov(\xi_k, \theta_l) = 0$, for all $m = 1, \ldots, M$ and $l = 1, \ldots, P$, then unconditional stochastic independence $\boldsymbol{Y} \perp \boldsymbol{D}$ is implied. In this case, the missing data mechanism is MCAR. In other words, if $\Sigma_{\boldsymbol{\xi},\boldsymbol{\theta}}$ is block diagonal, so that

$$\Sigma_{\boldsymbol{\xi},\boldsymbol{\theta}} = \begin{pmatrix} \Sigma_{\boldsymbol{\xi}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\boldsymbol{\theta}} \end{pmatrix}, \tag{4.84}$$

then stochastic independence $\boldsymbol{Y} \perp \boldsymbol{D}$ follows, indicating that the nonresponse mechanism is MCAR.

This can also be shown by considering the likelihood function, which is generally given by

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} P(D_{ni} = d_{ni} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}). \tag{4.85}$$

Using MML estimation, that is,

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} \int_{\mathbb{R}^p} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} P(D_{ni} = d_{ni} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}) g(\boldsymbol{\xi}, \boldsymbol{\theta}), \tag{4.86}$$

which follows from of Equation 4.76 taking the assumption of conditional stochastic independence $D_i \perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}$ into account. Given $\Sigma_{\boldsymbol{\xi},\boldsymbol{\theta}}$ is block diagonal since $\boldsymbol{\xi} \perp \boldsymbol{\theta}$, Equation 4.85 becomes

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} \int_{\mathbb{R}^p} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} P(D_{ni} = d_{ni} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}) g(\boldsymbol{\xi}) g(\boldsymbol{\theta}), \tag{4.87}$$

since $g(\boldsymbol{\xi}, \boldsymbol{\theta}) = g(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}) g(\boldsymbol{\theta}) = g(\boldsymbol{\xi}) g(\boldsymbol{\theta})$. This allows to write Equation 4.87 as

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^m} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} g(\boldsymbol{\xi}) \prod_{n=1}^{N} \int_{\mathbb{R}^p} \prod_{i=1}^{I} P(D_{ni} = d_{ni} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}) g(\boldsymbol{\theta}). \tag{4.88}$$

Hence, the likelihood can be factorized into two independent pieces and $\boldsymbol{D}$ needs not to be modeled jointly with $\boldsymbol{Y}$ (see Section 4.5.1). The missing data mechanism is ignorable.

The variance-covariance matrix $\Sigma_{\xi,\theta}$ can be estimated using MML estimation. The correlations between the latent dimensions ξ_m and θ_l allow to examine and to quantify the strength of the dependencies between the occurrence of nonreponses and the latent proficiency of interest. Hence, MIRT models for nonignorable missing data and especially BMIRT models are of additional diagnostic value.

Application of the BMIRT model to Data Example A The BMIRT model was applied to Data Example A. Two models were estimated with the BMIRT Rasch (1PL-BMIRT) model using ConQuest (Wu et al., 1998) and the two-parameter BMIRT (2PL-BMIRT) model using Mplus 6 (Muthén & Muthén, 1998 - 2010). Data Example A was generated under the validity of the Rasch model. Hence, the choice of the 1PL-BMIRT model is adequate. In this case, all discrimination parameters in $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma_\theta}$ are fixed to zero or one. In Data Example A, ξ and θ were unidimensional each, implying that $\boldsymbol{\alpha}$, $\boldsymbol{\gamma_\theta}$ and $\boldsymbol{\Lambda}$ are identity matrices. ConQuest allows for estimation of ML, WML, and EAP person parameter estimates. Unfortunately, Mplus 6 allows only for EAP-person parameter estimation. The primary goal of applying the 2PL-BMIRT model to Data Example A is to study the affect of the model choice to item discrimination estimates compared to the model that ignores missing data. Generally, all item and person parameter estimates of the 1PL- and 2PL-BMIRT models were compared with the true values, the estimates obtained from the complete data using the undimensional IRT model of ξ based on $\boldsymbol{Y}$, and the estimates obtained from incomplete data using the undimensional IRT model of ξ based on $\boldsymbol{Y}$, which ignores missing responses.

At first, the estimated **item difficulties** are considered. The left graph of Figure 4.15 shows the estimated item difficulties obtained by the BMIRT Rasch model compared to the true parameters respectively. Additionally, Table 4.9 gives the $\hat{\beta}_i$ from different models, including those of the BMIRT Rasch model. The mean bias of the 30 difficulty estimates was 0.035. This is not significantly different from zero ($t = 1.564$, $df = 29$, $p = 0.129$). Recall that the mean bias of the estimated item difficulties in the unidimensional IRT model that ignores missing data was significantly negative ($Bias = -0.076$, $t = -2.868$, $df = 29$, $p = 0.008$). The bias reduction is also reflected by the MSE which is 0.016 in the BMIRT Rasch model instead of 0.026 when missing data were ignored. The slope of the regression of the estimates $\hat{\beta}_i$ on the true values $\beta_i$was not significantly different from one (Slope $= 0.981$, $SE = 0.017$ $t = -1.130$, $p = 0.461$). Hence, the

remaining bias in the BMIRT Rasch model is unsystematic with respect to the estimands β_i although the item difficulties are strongly correlated with the the probability of item nonresponse ($P(D_i = 0)$). In contrast in the unidimensional model ignoring missing data a systematic bias was found indicated by a slope significantly different from one (Slope = 0.938, $SE = 0.017$ $t = -3.700$, $p = 0.002$). The reason is that more difficult items were more likely answered by, on average, more proficient persons feigning easier items (see Section 3.2.2). The 1PL-BMIRT model corrects for the systematic missing responses of difficult items by less proficient persons.

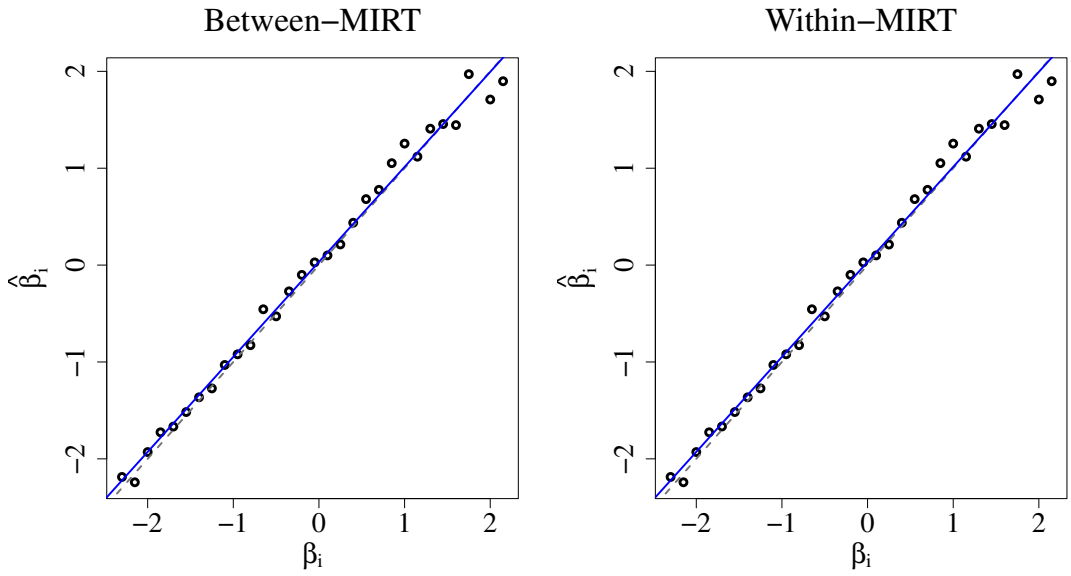

Figure 4.15: Comparison of true and estimated item difficulties of the BMIRT Rasch model (left) and the W_{Dif}MIRT Rasch model (right) for nonignorable missing data (Data Example A). The grey dotted line is the bisectric. The blue line is the regression line.

Using Mplus 6, the **item discrimination** parameters were freely estimated in the 2PL-BMIRT model. Data Example A was generated using the Rasch model for the items Y_i and the response indicators D_i. However, in real applications a researcher does not know the true data generating model and might favor the 2PLM. Furthermore, here the estimates $\hat{\alpha}_i$ were compared between the 2PL-BMIRT model and the unidimensional model that ignores missing data. For identification of the 2PL-BMIRT, the latent distributions were fixed to $E(\xi) = E(\theta) = 0$ and $Var(\xi) = Var(\theta) = 1$. All parameters α_i and $\gamma_{i;\theta}$ were freely estimated. Figure 4.16 shows the discrimination estimates of the 30 items of Data Example A obtained with the 2PL-BMIRT model (left). Compared with the discrimi-

Table 4.9: True Parameters β_i and Estimates $\hat{\beta}_i$ and $\hat{\gamma}_{i0}$ for Different Models: the Unidimensional IRT Model With Complete Data and With Incomplete Data, the BMIRT Rasch, and the W_{Dif}MIRT Rasch Model.

	True	Complete	missing	BMIRT Rasch		W_{Dif}MIRT Rasch	
Item	β_i	$\hat{\beta}_i$	$\hat{\beta}_i$	$\hat{\beta}_i$	$\hat{\gamma}_{i0}$	$\hat{\beta}_i$	$\hat{\gamma}_{i0}$
1	-2.300	-2.258	-2.209	-2.187	-2.520	-2.187	-2.520
2	-2.150	-2.238	-2.262	-2.242	-2.411	-2.242	-2.411
3	-2.000	-1.952	-1.964	-1.930	-1.997	-1.930	-1.997
4	-1.850	-1.756	-1.770	-1.726	-1.840	-1.726	-1.840
5	-1.700	-1.693	-1.728	-1.667	-0.918	-1.666	-0.918
6	-1.550	-1.520	-1.574	-1.517	-1.508	-1.517	-1.508
7	-1.400	-1.349	-1.441	-1.365	-1.092	-1.365	-1.092
8	-1.250	-1.274	-1.327	-1.273	-1.367	-1.273	-1.367
9	-1.100	-1.024	-1.081	-1.031	-1.445	-1.031	-1.445
10	-0.950	-0.951	-0.981	-0.921	-1.373	-0.921	-1.373
11	-0.800	-0.836	-0.922	-0.827	-0.579	-0.827	-0.579
12	-0.650	-0.486	-0.529	-0.456	-1.077	-0.456	-1.077
13	-0.500	-0.511	-0.595	-0.529	-1.121	-0.529	-1.121
14	-0.350	-0.285	-0.376	-0.271	-0.509	-0.271	-0.509
15	-0.200	-0.072	-0.163	-0.101	-1.448	-0.101	-1.448
16	-0.050	-0.043	-0.058	0.028	-0.775	0.028	-0.775
17	0.100	0.100	-0.039	0.100	0.521	0.100	0.521
18	0.250	0.309	0.105	0.213	-0.100	0.212	-0.100
19	0.400	0.428	0.279	0.437	0.914	0.437	0.914
20	0.550	0.608	0.523	0.681	0.952	0.681	0.952
21	0.700	0.818	0.625	0.777	0.909	0.777	0.909
22	0.850	0.980	0.886	1.052	1.223	1.052	1.223
23	1.000	1.123	1.065	1.254	2.071	1.254	2.071
24	1.150	1.174	0.937	1.119	1.455	1.119	1.455
25	1.300	1.450	1.232	1.408	1.570	1.408	1.570
26	1.450	1.520	1.289	1.456	0.975	1.456	0.975
27	1.600	1.514	1.262	1.445	2.001	1.445	2.001
28	1.750	1.840	1.786	1.971	1.751	1.971	1.751
29	2.000	1.958	1.519	1.709	2.080	1.709	2.080
30	2.150	2.183	1.694	1.898	2.117	1.898	2.117
Mean Bias	-	0.043**	−0.076**	0.035	-	0.035	-
MSE	-	0.006	0.026	0.016	-	0.016	-

Note: * significant at 0.05 level (2-tailed); ** significant at 0.01 level (2-tailed).

nation estimates of the unidimensional 2PL model that ignores missing data, only small differences were found. The mean bias of $\hat{\alpha}_i$ obtained from the BMIRT model was 0.014, which is not significantly different from zero ($t = 0.636$, $df = 9$, $p = 0.530$). Recall that in the unidimensional model that ignores missingness the bias was also not different from zero (Bias = -0.019, $t = -0.888$, $df = 9$, $p = 0.382$). Similarly, the mean squared errors of the discrimination estimates of the two models were very close. In both models, the 2PL-BMIRT and the unidimensional model ignoring missing data, the *MSE* was about 0.014. Additionally, a correlation of $r = 0.937$ between the discrimination estimates of the unidimensional model that ignores missing data and the 2PL-BMIRT highlight the agreement of the results. In line with the findings of Rose et al. (2010), the item discrimination parameters turned out to be less affected by nonignorable missing data as well as choice of model to account for missing data. Hence, the application of the 2PL-BMIRT model hardly changed discrimination estimates.

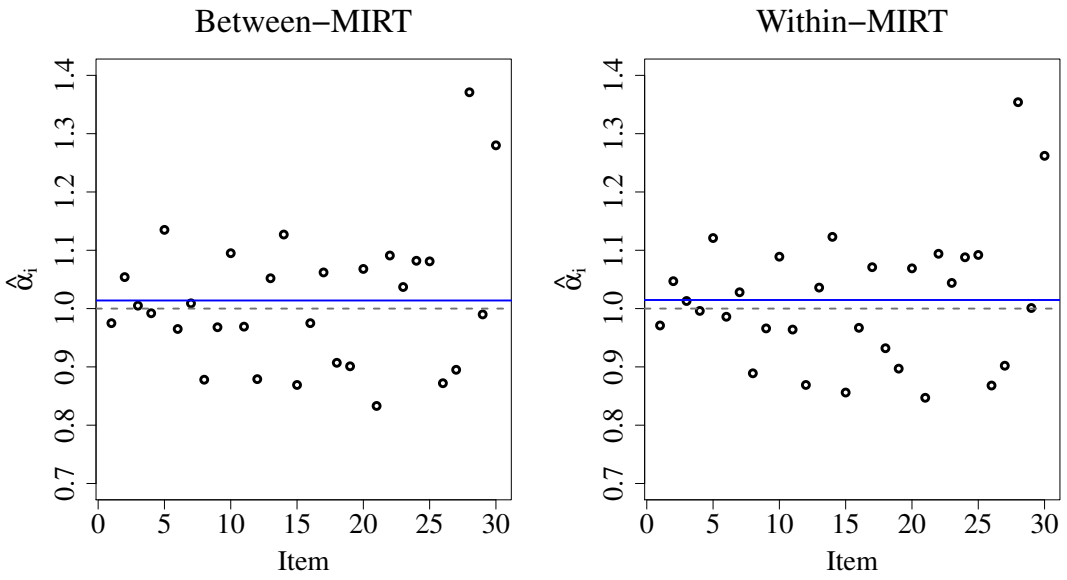

Figure 4.16: Item discrimination estimates of the 2PL-BMIRT model (left) and the 2PL-W$_{Dif}$MIRT model (right) for nonignorable missing data (Data Example A). The grey dotted line indicates the true value $\alpha_i = 1$ and the blue line indicates the mean $\bar{\hat{\alpha}}_i$.

Finally, the person parameter estimates were considered, starting with the **ML and WML person parameter estimates**. Table 4.10 shows some summary statistics of the ML-, WML-, and EAP-estimates obtained by the 1PLM-MIRT Rasch model. Due to

identification of the model, the mean is approximately zero. The variances of the ML- and WML-estimates are close to those of the unidimensional IRT model that ignores missing data (see Table 3.3). As Figures 4.17 and 4.18 confirm, ML and WML estimates of the BMIRT Rasch model and the unidimensional model ignoring missingness are almost identical. Accordingly, the correlations $r(\xi, \hat{\xi}_{ML}) = 0.819$ and $r(\xi, \hat{\xi}_{WML}) = 0.830$ in the BMIRT Rasch model are similar to $r(\xi, \hat{\xi}_{WML}) = 0.816$ and $r(\xi, \hat{\xi}_{WML}) = 0.827$ in the unidimensional model that ignores missing responses. Recall that in Section 3.1.3 it was demonstrated that the bias of ML- and WML-estimates depends strongly on the bias of the item parameter estimates especially the item difficulties. In Data Example A, the estimates $\hat{\beta}_i$ were only slightly biased when missing data were ignored. In conjunction with previous results of the simulation study presented in Chapter 3, the findings indicate that ML- and WML-estimates of the unidimensional model that ignores missingness and the BMIRT model for nonignorable missing data differs only when the item parameter estimates will be reasonably different. This implies that the accuracy of ML- and WML-

Table 4.10: Summary Information of ML-, WLM-, and EAP Person Parameter Estimates for the BMIRT Rasch Model for Nonignorable Missing Data (Data Example A).

Estimator	Mean	Variance	$r(\xi, \hat{\xi})$	$Rel(\hat{\xi})$	MSE	$r(bias, \xi)$
ML	0.052	1.610	0.819	0.673	0.540	0.052
WML	-0.001	1.401	0.830	0.650	0.438	-0.029
EAP	-0.001	0.759	0.883	0.771	0.222	-0.493

person parameter estimates cannot be increased by the BMIRT Rasch model. In fact, in Data Example A, the standard errors of the ML- and WML estimates of the BMIRT and the unidimensional model correlates approximately to one as well. In line with these findings, the marginal reliabilities $Rel(\hat{\xi}_{ML}) = 0.673$ and $Rel(\hat{\xi}_{WML}) = 0.650$ under the BMIRT Rasch model are close to $Rel(\hat{\xi}_{ML}) = 0.666$ and $Rel(\hat{\xi}_{WML}) = 0.641$ obtained by the unidimensional model that ignores missing data. The mean squared errors confirm that the bias reduction of the ML- and WML-person parameters was negligible in Data Example A. The reason is that information given by the individual missing pattern $\boldsymbol{D_n} = \boldsymbol{d_n}$ or the latent correlation $Cor(\xi, \theta)$ is not taken into account in ML and WML person parameter estimation. This is an important difference to Bayesian persson parameter estimation.

EAP-person parameter estimates are Bayesian estimates that have different properties than ML and WML person parameter estimates. As Figure 4.19 shows, the EAPs obtained under the unidimensional model ignoring missing data and the BMIRT Rasch model are different. Although still high, the correlation of the EAPs of both models is

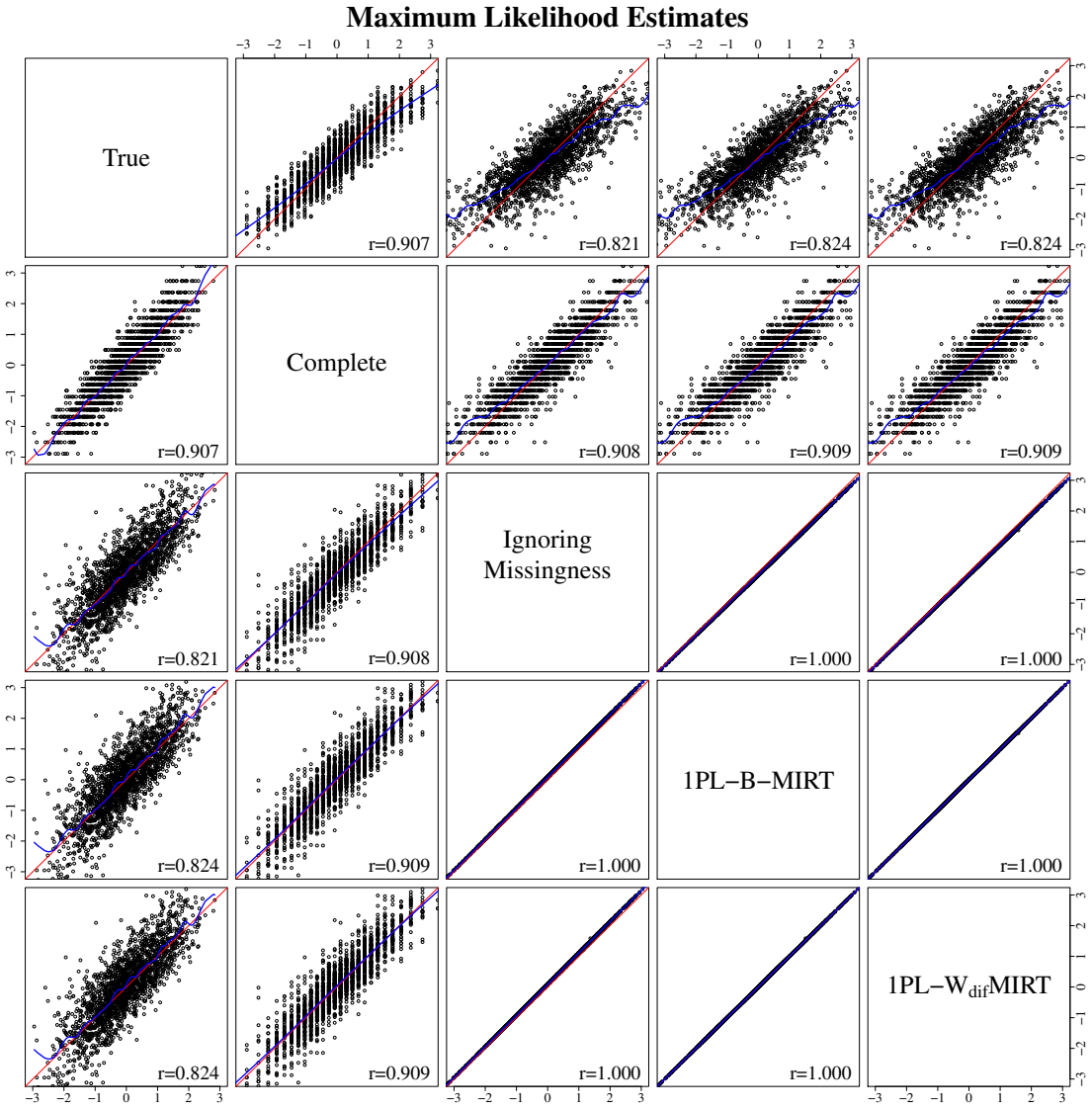

Figure 4.17: True values of ξ and ML person parameter estimates obtained by different IRT models (Data Example A). The red lines represent the bisectric. The blue lines are smoothing spline regressions.

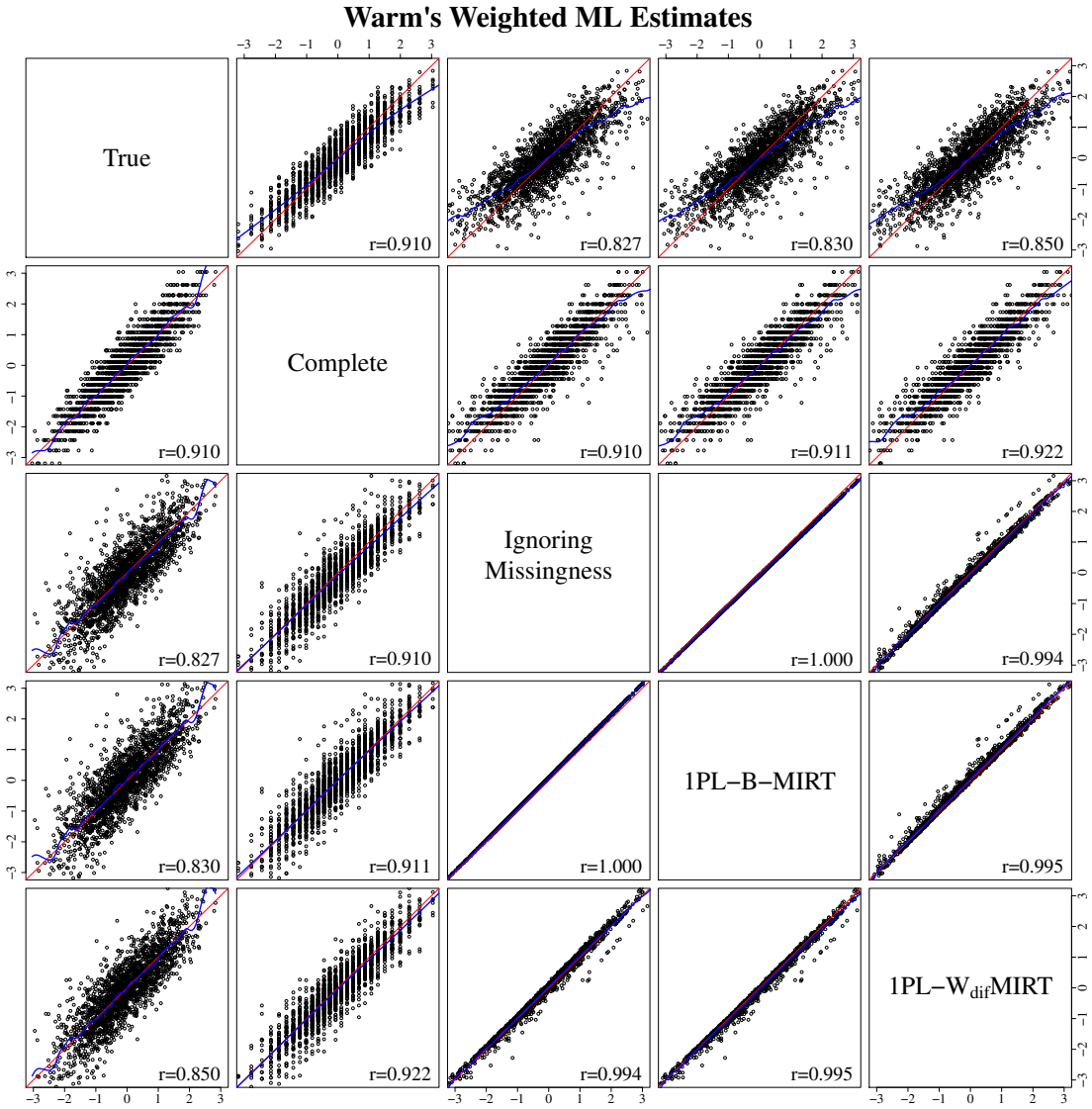

Figure 4.18: True values of ξ and Warm's weighted ML person parameter estimates obtained by different IRT models (Data Example A). The red lines represent the bisectric. The blue lines are smoothing spline regressions.

substantially lower than one ($r(\xi, \hat{\xi}_{EAP}) = 0.934$). The variances of the EAPs obtained from these two models differ as well. The variance is $Var(\hat{\xi}_{EAP}) = 0.759$ in the BMIRT Rasch model, compared to $Var(\hat{\xi}_{EAP}) = 0.632$ in the unidimensional model that ignores missing responses. Recall that the variance was $Var(\hat{\xi}_{EAP}) = 0.859$ when the complete data were used in the unidimensional model (cf. Table 3.3). Generally, the less information is available, the stronger the impact of the prior distribution on parameter estimation is, and, therefore, the stronger the shrinkage toward $E(\xi)$ is. Missing data means a loss of observed information with respect to the estimand ξ resulting in a substantial variance reduction compared to the complete data model. The BMIRT Rasch model as well as the 2PL-BMIRT model reduce the shrinkage of EAPs using the information of $\boldsymbol{D}$ with respect to ξ. As a result, the correlation between the latent variable ξ and the bias of the EAP-estimates reduces. In the BMIRT Rasch model it was $r(\xi, Bias_{EAP}) = -0.493$, compared to $r(\xi, Bias_{EAP}) = -0.608$ in the unidimensional model ignoring missing data. Finally, Table 4.10 reveals that the MSE of the EAP drops from $MSE(\hat{\xi}_{EAP}) = 0.327$ when the missing data were ignored to $MSE(\hat{\xi}_{EAP}) = 0.222$ in the BMIRT Rasch model. The accuracy of the EAPs was reasonably improved in the BMIRT Rasch model.

How does EAP person parameter estimation use the information about the latent ability of interest that is given by missingness? For the estimation of individual values of a single latent dimension in a multidimensional latent trait model, ML and WML estimators use only the information from that manifest variables Y_i that directly indicate the measurement model of this latent dimension. Latent covariances as well as information of covariates in the background model are not used for person parameter estimation. In the case missing data, that means that only the observed response vector $\boldsymbol{Y}_{n;obs} = \boldsymbol{y}_{n;obs}$ is used for estimation of individual values of $\boldsymbol{\xi}$ of each case n. In contrast, EAP estimation of person parameters accounts for the latent covariances $Cov(\xi_m, \theta_l)$ by the prior distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ that is involved in the estimation procedure. Furthermore, all information given by the observed data $(\boldsymbol{Y}_{n;obs}, \boldsymbol{D}_n) = (\boldsymbol{y}_{n;obs}, \boldsymbol{d}_n)$ is exploited for estimation of persons' latent ability $\boldsymbol{\xi}$. The gain of information due to manifest variables that are not direct indicators of a latent dimension increases, the higher the correlations between the latent variables are. For that reason, EAP and MAP estimates are typically preferred in multidimensional adaptive testings with correlated latent abilities (Segall, 1996, 2000; Frey & Seitz, 2009). For deeper understanding, the estimation equation of $\hat{\xi}_{EAP}$ is considered. In Data Example

A, ξ and θ are each unidimensional. In this case, the EAP of the latent ability is given by

$$\hat{\xi}_{EAP} = \frac{\int_{\mathbb{R}} \xi \cdot \int_{\mathbb{R}} g(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \,|\, \xi; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \theta; \boldsymbol{\phi}) g(\xi, \theta) d\xi d\theta}{\int_{\mathbb{R}} \int_{\mathbb{R}} g(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \,|\, \xi; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \theta; \boldsymbol{\phi}) g(\xi, \theta) d\xi d\theta}. \tag{4.89}$$

The joint distribution of the latent variables in the nominator of Equation 4.89 can be factorized, so that $g(\xi, \theta) = g(\xi \,|\, \theta) g(\theta)$. Hence, different values of ξ are more or less likely, given the values of θ. In most applications, a bivariate normal distribution is assumed for $g(\xi, \theta)$, which is sufficiently described by the vector of expected values - here $E(\xi)$ and $E(\theta)$ - and the covariance matrix $\Sigma_{\xi,\theta}$. If ξ and θ are linearly regressively dependent ,then the conditional distribution $g(\xi \,|\, \theta)$ in the normal model can be characterized by a linear regression $E(\xi \,|\, \theta)$ and a normally distributed residual $\varepsilon_\xi = \xi - E(\xi \,|\, \theta)$. In terms of probability, the more a value of ξ deviate from the expected values $E(\xi \,|\, \theta = \theta)$, the less likely this value and more extreme values are. In fact, if θ would be known, then the ability estimates would not shrink toward the mean $E(\xi)$ but to the individual conditional expected values $E(\xi \,|\, \theta = \theta)$[16]. Nevertheless, the shrinkage effect due to item nonresponses can be considerably reduced if ξ and θ are reasonably correlated and if θ can be reliably estimated based on $\boldsymbol{D}$. This was also found in Data Example A. Table 4.10 shows that the variance of the EAPs is $Var(\hat{\xi}_{EAP}) = 0.759$. In the unidimensional model that ignores missing data, the variance was $Var(\hat{\xi}_{EAP}) = 0.632$ (see Table 3.3). The increas in the variance of the EAPs marks the reduced shrinkage effect and, therefore, an increased reliability $Rel(\hat{\xi}_{EAP}) = 0.771$. As Figure 4.19 illustrates, the EAPs of the unidimensional model that ignores missing data and the BMIRT Rasch model are different. A careful inspection reveals that the EAP estimates are especially downward corrected in cases with below-average proficiency levels where the proportions of missing data were on average higher.

These findings can be generalized to cases with m-dimensional latent abilities $\boldsymbol{\xi}$ and P-dimensional latent propensities $\boldsymbol{\theta}$. With MML estimation, a multivariate normal distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ is assumed inmost applications. The covariance matrix $\Sigma_{\boldsymbol{\xi},\boldsymbol{\theta}}$ describes the mutual linear relations between all latent variables ξ_m and θ_l. In this case, information from all other latent dimensions θ_l and $\xi_{k \neq m}$ is taken into account for EAP estimation of a

[16] In application θ is typically unknown as well and the estimates $\hat{\theta}_{EAP}$ shrink in turn to $E(\theta \,|\, \xi = \xi)$. As a consequence, there is a shrinkage of $(\hat{\xi}_{EAP}, \hat{\theta}_{EAP})$ toward the vector of expected values $E(\xi)$ and $E(\theta)$.

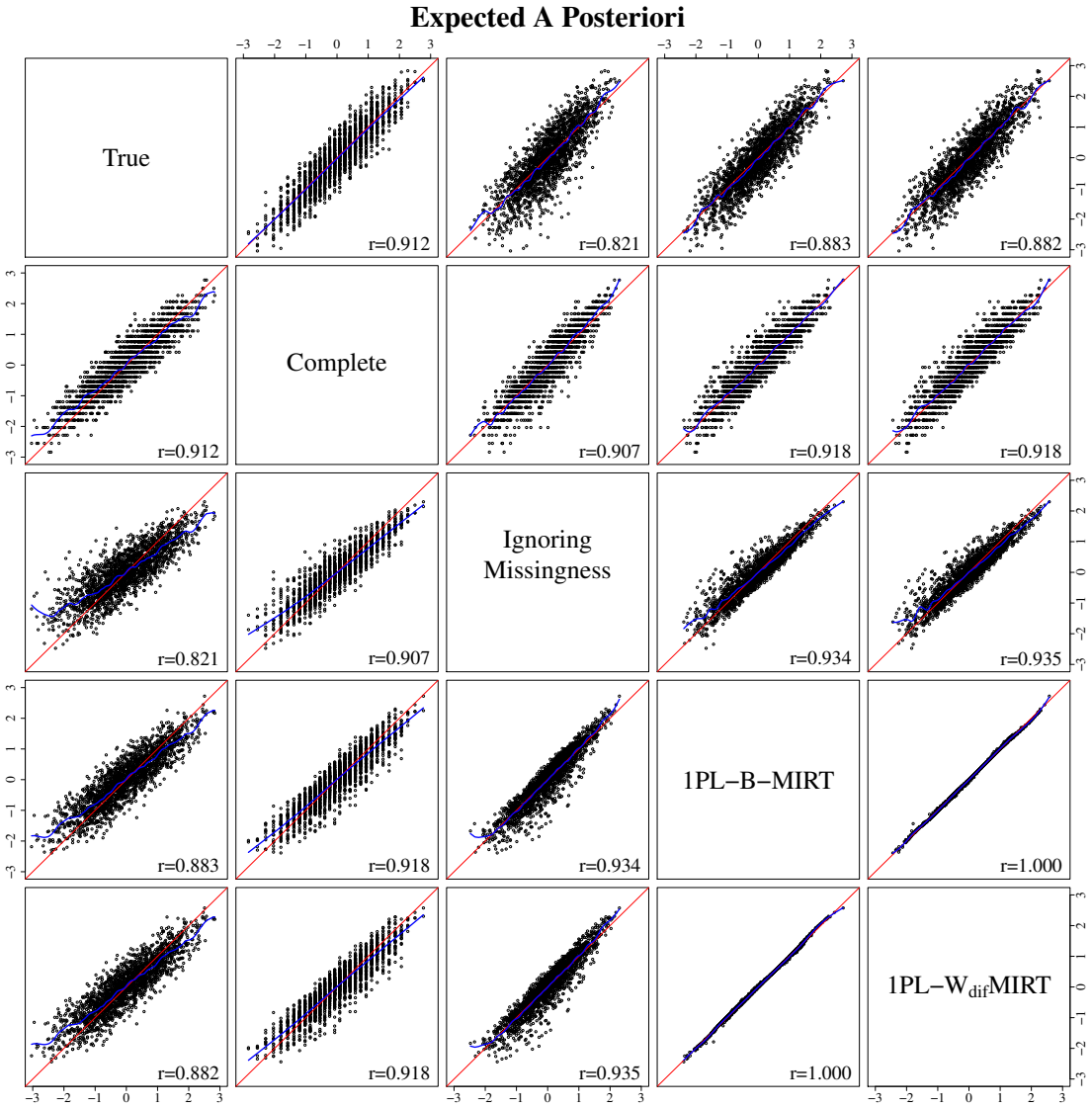

Figure 4.19: True values of ξ and EAP person parameter estimates obtained by different IRT models (Data Example A). The red lines represent the bisectric. The blue lines are smoothing spline regressions.

single dimension ξ_m. Equation 4.89 can be generalized to

$$\hat{\xi}_{m;EAP} = \frac{\int_{\mathbb{R}} \xi_m \cdot \int_{\mathbb{R}^{m-1}} \int_{\mathbb{R}^p} P(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \,|\, \boldsymbol{\xi}) P(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{\theta}) g(\boldsymbol{\xi}, \boldsymbol{\theta}) d\boldsymbol{\xi} d\boldsymbol{\theta}}{\int_{\mathbb{R}^m} \int_{\mathbb{R}^p} P(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \,|\, \boldsymbol{\xi}) P(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{\theta}) g(\boldsymbol{\xi}, \boldsymbol{\theta}) d\boldsymbol{\xi} d\boldsymbol{\theta}}. \quad (4.90)$$

This implies that not only information of missingness is used, but also information from all other ability dimensions ξ_k that are correlated with ξ_m. Furthermore, manifest covariates $\boldsymbol{Z} = Z_1, \ldots, Z_J$ that are predictive with respect to the latent dimensions ξ_m or which are informative with respect to missingness can also be included in a latent regression model with $E(\boldsymbol{\xi} \,|\, \boldsymbol{Z})$ and $E(\boldsymbol{\theta} \,|\, \boldsymbol{Z})$. In this case, the prior distribution used for EAP estimation in Equation 4.90 is replaced by the conditional distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{Z})$. Informative covariates are useful twofold: (a) They reduce the shrinkage effect, and (b) they can improve parameter estimation (Mislevy, 1987, 1988).

In summary, the BMIRT model was derived as an example of MIRT models for nonigorable missing responses. Applied to Data Example A, the systematic bias of item difficulties caused by nonignorable missing responses was removed. Item discrimination estimates were found to be unbiased when missing responses are ignored. Hence, with the 2PL-BMIRT model, similar discrimination estimates were obtained. The three different person parameter estimates - ML, WML, and EAP - showed considerable differences. ML and WML estimation of latent dimensions ξ_m depends only on item parameter estimates and item responses to those items Y_i that directly indicate ξ_m. Neither responses to other indicators Y_j not indicating ξ_m, nor correlations of latent dimensions $Cor(\xi_m, \theta_l)$, nor informative background variables $\boldsymbol{Z}$ have any affect on ML and WML estimation. All corrections of these estimates in the BMIRT model is a result of corrected item parameters in the measurement model of $\boldsymbol{\xi}$. In light of these findings, Bayesian estimates such as EAPs may be superior to ML and Warm's weighted ML person parameter estimates, since additional diagnostic information is utilized for ability estimation[17]. Most importantly, EAP estimation includes prior information given by the distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ of the latent ability and the latent response propensity. The latter is indicated by the missing indicator vector $\boldsymbol{D}$. In this way, the information of missingness is used for person parameter estimation. In conjunction with the results of the simulation study in Chapter 3, it can be concluded that ML and WML person parameter estimates obtained by the MIRT model for nonignorable missing data only differ compared to that of the model that ignores missing data when the

[17] The same is true for Maximum A Posteriori (MAP) estimates. MAPs were not examined here. However, EAPs and MAPs rest upon the same individual posterior distributions and have, therefore, very similar properties.

item parameters in the measurement model of $\boldsymbol{Y}$ differ. This implies that ML and WML estimators of $\boldsymbol{\xi}$ do not make use of additional information provided by the missing pattern $\boldsymbol{D}$. Bayesian estimators such as the EAP uses this information by integrating over the joint distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ of the latent variables.

4.5.3.3 Within-item Multidimensional IRT Models for Nonignorable Missing Data

Within item dimensional models have become popular for scaling tests consisting of items that require more than one latent ability to provide a correct response (e. g. Ackerman, 1994; Ackerman, Gierl, & Walker, 2003; Hartig & Höhler, 2009; Reckase, 1985; Wang et al., 1997). Hence, in these models, stochastic dependencies of single items on more than one latent variable can be modeled appropriately. Especially in cognitive psychology, the process of solving an item may depend on several skills. Furthermore, IRT models for within-item multidimensional items are useful for applications to repeated measurements. Items repeatedly presented to test takers can be assumed to be stochastically dependent on: (a) the initial ability level at the first measurement occasion and (b) the change in the latent ability that potentially has taken place between the first and subsequent measurement occasions. Hence, latent change IRT models can also be regarded as within-item multidimensional IRT models (Embretson, 1991; Meiser, 2007; von Davier, Xu, & Carstensen, 2011). There might be many other applications where it is theoretically required to model stochastic dependencies of an item with more than one latent dimension.

Within-item multidimensional MIRT models have also been proposed as alternative models for BMIRT models for nonignorable missing responses (Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). Typically, BMIRT and WMIRT models for missing responses are considered to be equivalent. Indeed, for the case of the Rasch model, Rose et al. (2010) demonstrated that both models - the 1PL-BMIRT and the 1PL-WMIRT model - are equal in terms of model fit and with respect to the model parameters $\boldsymbol{\iota}$. Hence, the item difficulties β_i are equal and the person variable $\boldsymbol{\xi}$ is equivalently constructed in both models. However, the item parameters referring to the measurement model based on $\boldsymbol{D}$ as well as the meaning of the latent variable $\boldsymbol{\theta}$ changes fundamentally. In general, the interpretability of item and person parameters in WMIRT models can become challenging. Recall that latent variables are constructed in a measurement model. Accordingly, the meaning and the interpretation of parameters depends strongly on the model specification. Hitherto, barely much attention was paid to this fact. This is all the more remarkable as several competing WMIRT models for nonignorable missing data can be derived. Can all theses model be used interchangeably?

Are all of these models equally suited to account for missing responses? To answer these questions, the different 1PL- and 2PL-WMIRT models will be derived step by step. It will be shown that the applicability of a particular WMIRT model introduced below depends a priori on the decision for either the 1PL- or the 2PL-MIRT model. Note that the decision for using the Rasch model or the Birnbaum model is often made in the run-up of educational and psychological testings. This decision may limit the range of applicable WMIRT models for item nonresponses. For that reason, the different WMIRT models will be derived separately for the 1PLM and the 2PLM. The issue of model equivalence is explicitly taken into account in the derivations of the different WMIRT models.

Model equivalence in MIRT models for nonignorable missing data The issue of model equivalence was repeatedly addressed in SEM (e. g. Raykov & Penev, 1999; Raykov & Marcoulides, 2001). Typically, measurement models are considered to be equivalent if they have the same model fit and, therefore, the same statistical fit indexes (Raykov & Marcoulides, 2001). However, as Raykov and Marcoulides emphasized, the substantial meaning of two equivalent models can be very different. Considering that MIRT models for item nonresponses should correct for missing responses without altering the meaning of the latent ability variable $\boldsymbol{\xi}$ and the model parameters $\boldsymbol{\iota}$, the term model equivalence is used here in a stricter sense. Let there be two models: Model A and Model B. Both can be equivalent with respect to three criteria:

1. The latent ability variables in Models A and B are constructed in exactly the same way as in the target model, which is the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$.

2. The bias of item and person parameters due to missing responses is equally reduced in both models.

3. Both models fit given empirical data equally well.

The first criterion is essential. If $\boldsymbol{\xi}$ is not identically constructed, then the models do not simply correct for item nonresponses but consist of parameters with a different meaning. If Model A, Model B, or both are not equivalent to the target model they cannot be used to correct for item nonresponses. Even if A and B are equivalent in the construction of $\boldsymbol{\xi}$, they may differ regarding the reduction of the missing-induced bias. In this case the two models are not equivalent in terms of bias adjustment, indicating that one model is superior to the other model and should be preferred in application. The third criterion, the equivalence of model fit, is the least important criterion, which can be used for model

diagnosis. Many different measures have been proposed to quantify the fit of models to observed data. Such fit indices typically rest upon two pieces of information: (a) the discrepancy between observed data and expected data (residuals) given the sample estimates, and (b) model complexity. Hence, if A and B are not equivalent in terms of model fit, this indicates that one of the models is superior to the other model in terms of lower residuals and/or parsimony. However, if all three criterions are fullfilled, Models A and B can be used interchangeably. Both imply the same joint distribution $g(\boldsymbol{Y}, \boldsymbol{D}; \boldsymbol{\iota}, \boldsymbol{\phi})$ with $\boldsymbol{\iota}$ equal in both models. However, $\boldsymbol{\phi}$ can be different.

In the following, it will be demonstrated that at least two WMIRT models can be derived that are equivalent to the BMIRT model introduced above. In the first model, denoted by W_{Dif}MIRT, a potentially mulidimensional latent difference variable $\boldsymbol{\theta}^*$ is defined. In the second model, the W_{Res}MIRT model, $\tilde{\boldsymbol{\theta}}$ is constructed as a latent residual. In both approaches, the construction of $\boldsymbol{\xi}$ is unchanged and the parameter vector $\boldsymbol{\iota}$ remains unaffected. Hence, the target measurement model is preserved in the joint model based on $(\boldsymbol{Y}, \boldsymbol{D})$. It will be studied whether the bias due to nonignorable missing data is equally reduced by the different models. Furthermore, the applicability of the alternative models will be examined. At first, the WMIRT Rasch (1PL-WMIRT) models are derived and applied to Data Example A. The 2PL-WMIRT models will be developed and demonstrated afterwards.

Within item multidimensional Rasch model The WMIRT Rasch model requires that the conditional independence assumptions given by Equations 4.5.4 and 4.74 hold. Especially, the conditional stochastic independence $Y_i \perp (\boldsymbol{Y}_{-i}, \boldsymbol{D}) \,|\, \boldsymbol{\xi}$ is essential to ensure equivalence in the construction of the latent variable $\boldsymbol{\xi}$. The second assumption $D_i \perp (\boldsymbol{D}_{-i}, \boldsymbol{Y}) \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$ allows the response indicators D_i not only to be stochastically dependent on $\boldsymbol{\xi}$. Accordingly, the general model equation of the logits (see Equation 4.80) allows the discrimination parameters $\boldsymbol{\gamma}_{\xi}$ to be different from zero. In that case, the response indicators are conditionally stochastically dependent on $\boldsymbol{\xi}$ given $\boldsymbol{\theta}^*$[18]. This is the distinctive characteristic of all WMIRT models described here. Recall that in BMIRT models $\boldsymbol{D} \perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}$ follows from Equation 4.81. Note that the latent variables $\boldsymbol{\theta}^*$ or $\tilde{\boldsymbol{\theta}}$ in the WMIRT models are marked by the symbols $*$ or $\sim$. Similarly, some model parameters, such as γ^*_{ξ} or $\tilde{\gamma}^*_{\xi}$, are flagged with these symbols. This notation is used to highlight that alternative WMIRT models are specified differently, which result in a different construction of latent variables. Whereas the latent variable $\boldsymbol{\xi}$ is constructed equivalently in all

[18] Strictly speaking, $\boldsymbol{D} \not\perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}^*$ is implied if $\boldsymbol{\gamma}_{\xi} \neq \mathbf{0}$ *and* $\boldsymbol{\gamma}_{\theta} \neq \mathbf{0}$

MIRT models, $\boldsymbol{\theta}$ can only be interpreted as a latent response propensity in BMIRT models. Hence, in within-item mulidimensional IRT models for nonignorale missing data, conditional stochastic dependency $\boldsymbol{D} \not\perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}^*$ or $\boldsymbol{D} \not\perp \boldsymbol{\xi} \,|\, \tilde{\boldsymbol{\theta}}$ is modeled, with $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$ as latent variables different from $\boldsymbol{\theta}$ of the BMIRT model. Rose et al. (2010) derived the WMIRT Rasch model rationally starting form the BMIRT Rasch model for the case of unidimensional variables ξ and θ. They demonstrated that in an equivalent WMIRT model, θ^* is constructed as a latent difference variable $\theta - \xi$. Accordingly, this model is denoted as W_{Dif}MIRT Rasch model or the 1PL-W_{Dif}model. The derivations of that model given by Rose et al. are briefly described here. Subsequently, the model will be generalized to the case of m-dimensional latent abilities $\boldsymbol{\xi}$ and p-dimensional latent variables $\boldsymbol{\theta}^*$.

In the 1PL-W_{Dif}MIRT model with unidimensional variables ξ and θ, the logits $l(Y_i)$ and $l(D_i)$ of the items Y_i and the response indicators D_i are

$$l(Y_i) = \xi - \beta_i \tag{4.91}$$

$$l(D_i) = \theta - \gamma_{i0} \tag{4.92}$$

Solving for the latent variables gives simply

$$\xi = l(Y_i) + \beta_i \tag{4.93}$$

$$\theta = l(D_i) + \gamma_{i0}. \tag{4.94}$$

Hence, the latent variables are the logits plus a constant given by the item difficulty or the threshold of the manifest variables Y_i and D_i. Due to model equivalence with respect to the construction of the latent ability ξ, Equations 4.91 and 4.93 applies also to the W_{Dif}MIRT model. The model equations of the logits $l(D_i)$, however, differ between the 1PL-BMIRT and 1PL-W_{Dif}MIRT model. In the latter, that is,

$$l(D_i) = \theta^* + \xi - \gamma_{i0}. \tag{4.95}$$

It is important to note that the logits $l(D_i)$ the 1PL-BMIRT and the 1PL-W_{Dif}MIRT model are equal. The person's log-odd to respond to an item i does not change by the choice of the model. Due to this equality, the right hand side of Equation 4.92 from the 1PL-BMIRT model can be inserted into Equation 4.95 yielding

$$\theta - \gamma_{i0} = \theta^* + \xi - \gamma_{i0}. \tag{4.96}$$

Solving for θ^* and rearranging gives

$$\begin{aligned} \theta^* &= \theta - \xi - \gamma_{i0} + \gamma_{i0} \\ &= \theta - \xi. \end{aligned} \tag{4.97}$$

θ^* is not a latent response propensity but a function $f(\xi, \theta)$ of the latent response propensity and the latent ability. More specifically, θ^* is constructed as a latent difference variable. Inserting Equations 4.93 and 4.94 into Equation 4.97 gives

$$\theta^* = l(D_i) + \gamma_{i0} - (l(Y_i) + \beta_i) \tag{4.98}$$

$$= l(D_i) - l(Y_i) + \gamma_{i0} - \beta_i. \tag{4.99}$$

Thus, in the two-dimensional W_{Dif}MIRT Rasch model, θ^* is a latent difference variable of the logits $l(D_i)$ and $l(Y_i)$ plus the constant $\gamma_{i0} - \beta_i$. The interpretation of some parameters in the model is more difficult compared to the 1PL-BMIRT model. If the correlation $Cor(\xi, \theta)$ is positive, then the correlation $Cor(\xi, \theta^*)$ in the 1PL-W_{Dif}MIRT model is usually negative. Information about the strength of the relationship between the tendency to respond to the test items and the latent ability is not directly given in the W_{Dif}MIRT Rasch model.

Application of the 1PL-W_{Dif}MIRT model to Data Example A The BMIRT Rasch model was identified by the restriction $E(\xi) = E(\theta) = 0$, while all parameters β_i and γ_{i0} were freely estimated. Similarly, the W_{Dif}MIRT Rasch model was identified by $E(\xi) = E(\theta^*) = 0$. ConQuest (Wu et al., 1998) was used for parameter estimation. A comparison of the parameter estimates of the 1PL-BMIRT and the 1PL-W_{Dif}MIRT model shows that item and person parameter estimates of both models are practically the same. The item difficulties β_i and the respective estimates are given in Table 4.9. Furthermore, Figure 4.15 illustrates the approximate equality of the estimates $\hat{\beta}_i$. Accordingly, the MSE = 0.016 of the difficulty estimates in the 1PL-W_{Dif}MIRT model was the same as in the 1PL-BMIRT. Similarly, ML, WML, and EAP person parameter estimates were nearly identical between the BMIRT and the W_{Dif}MIRT Rasch model. This can be seen in Figures 4.17 - 4.19. Small but negligible differences between estimates of the two models were only found in WML estimates. For that reason, a detailed description of the results is not repeated here. From the results, it is concluded that the reduction of the missing-induced bias in the 1PL-W_{Dif}model is the same as in the BMIRT Rasch model. The value of the log-likelihood of the 1PL-W_{Dif}model and the BMIRT Rasch was -46535.706. Since the

number of parameters was also equal ($n_{par} = 63$), the BIC of both models was identical as well (BIC = 93550.267; see Table 4.11). The results confirm that the two models are equivalent with respect to three criteria introduced previously: (1) the construction of $\boldsymbol{\xi}$, (2) the adjustment for nonignorable missing responses, and (3) the model fit.

Extending the 1PL-W$_{Dif}$MIRT model to multidimensional variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ Compared to the BMIRT Rasch model, not only the interpretability of some model parameters model parameters but also the model specification becomes increasingly challenging if the number of latent dimensions rises. This problem is exemplified here using the model that is graphically represented in Figure 4.14. If all non-zero item discriminations in this model are $\alpha_{im} = \gamma_{il} = 1$, then a four-dimensional 1PL-BMIRT model results. The specification of an equivalent W$_{Dif}$MIRT model is intricate in this example. One problem is that the factorial structure of $\boldsymbol{\theta}$ underlying $\boldsymbol{D}$ does not mirror the factorial structure of $\boldsymbol{\xi}$ underlying $\boldsymbol{Y}$. It might be intuitive that the response indicators of those items that constitute a distinct latent dimension ξ_m establish a distinct latent response propensity dimension θ_l as well. However, this is an assumption that does not need to hold in application. There might be other characteristics of the item which also determine the probability of a response, such as the response format. As Rose et al. (2010) found, items with open or constructed responses are generally more likely omitted than multiple choice items. If such item characteristics, which are independent of the item content, interact with person characteristics, then a complex multidimensional structure of $\boldsymbol{\theta}$ can result, which is potentially quite different from that of $\boldsymbol{\xi}$. Here it is argued that such a situation is very likely in real applications. Therefore, the BMIRT and WMIRT models will be generalized to cases with a multidimensional latent variable $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. Hence, the crucial question is: How does one specify an equivalent W$_{Dif}$MIRT model for non-ignorable missing data in general? Again, the term equivalent refers to three aspects: (1) $\boldsymbol{\xi}$ is constructed as in the complete data target model of $\boldsymbol{Y}$, (2) the adjustments of the item and person parameter estimates for missing responses is identical, and (3) the goodness-of-fit is equivalent. For the case of the 1PL-BMIRT and the 1PL-W$_{Dif}$MIRT models, this was easy to show when θ and ξ were each unidimensional. However, the idea of constructing θ^* as a difference $\xi - \theta$ (see Equation 4.97) needs to be adapted to cases with multidimensional latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. The idea presented here is to define a p-dimensional variable $\boldsymbol{\theta^*} = \theta_1^*, \ldots, \theta_P^*$ with each dimension defined as the difference

$$\theta_l^* \quad = \quad \theta_l - \sum_{m=1}^{M} \xi_m. \tag{4.100}$$

θ_l refers to the l-th dimension of $\boldsymbol{\theta}$ as defined in the equivalent BMIRT Rasch model. It was shown that $\theta_l = l(D_i) + \gamma_{i0}$. Inserting this expression in Equation 4.100 gives

$$\theta_l^* \quad = \quad l(D_i) + \gamma_{i0} - \sum_{m=1}^{M} \xi_m. \tag{4.101}$$

Thus, θ_l^* is a difference of the logit $l(D_i)$ and the sum of the latent ability dimensions ξ_m. In order to obtain the specification rules of the general 1PL-W$_{Dif}$MIRT model for multidimensional latent variables, Equation 4.100 needs to be solved for $l(D_i)$, yielding

$$l(D_i) \quad = \quad \theta_l^* + \sum_{m=1}^{M} \xi_m - \gamma_{i0}. \tag{4.102}$$

In general, the logit of a within-item dimensional item in the 1PLM is the weighted sum of the latent variables. The weights are the item discriminations that can only be zero or one in this model. Hence, they serve as indicator variables determining whether or not a particular item is conditionally stochastically dependent on a certain latent dimension. Accordingly, the logit $l(D_i)$ of each response indicator is modeled as the weighted sum of all M latent dimensions ξ_m and the latent difference variable θ_l^*. The resulting model equation for the response indicators D_i is

$$P(D_i = 1 \,|\, \boldsymbol{\xi}, \theta_l^*) \quad = \quad \frac{\exp(\theta_l^* + \sum_{m=1}^{M} \xi_m - \gamma_{i0})}{1 + \exp(\theta_l^* + \sum_{m=1}^{M} \xi_m - \gamma_{i0})}. \tag{4.103}$$

Since the construction of $\boldsymbol{\xi}$ needs to be unaffected by the choice of a particular model, the model equations for the items Y_i remain the same as in the complete data model of $\boldsymbol{Y}$ as well as the BMIRT and the W$_{Dif}$MIRT Rasch model for nonignorable missing data (Equation 4.77). Finally, the model equation of the complete vector of logits $\boldsymbol{l(Y, D)}$ can be written as

$$\begin{pmatrix} \boldsymbol{l(Y)} \\ \boldsymbol{l(D)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \mathbf{1} & \boldsymbol{\gamma_\theta} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\theta} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma_0} \end{pmatrix}. \tag{4.104}$$

Hence, all elements of the $(I \times M)$-dimensional sub-matrix $\boldsymbol{\gamma}_{\boldsymbol{\xi}}^*$ of $\boldsymbol{\Lambda}$ (cf. Equation 4.80) are $\gamma_{im}^* = 1$. The asterisk „$*$“ is used to differentiate the parameters of the W$_{Dif}$MIRT model from that of the BMIRT model. In other words $\boldsymbol{\gamma}_{\boldsymbol{\xi}} = \mathbf{0}$ designates the BMIRT Rasch model and $\boldsymbol{\gamma}_{\boldsymbol{\xi}}^* = \mathbf{1}$ the 1PL-W$_{Dif}$MIRT model. The sub-matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma_\theta}$ are equal in both models and do not need to be distinguished.

Returning to the hypothetical example presented in Figure 4.14, the 1PL-W$_{Dif}$MIRT Rasch model, which is equivalent to a 1PL-BMIRT Rasch model, is graphically depicted in Figure 4.20. Except for the latent covariances, all drawn paths are fixed to be one. In both the BMIRT and the W$_{Dif}$MIRT Rasch model, the elements of Λ are not estimable parameters but are fixed to zero or one in advance. Note that this model is generally

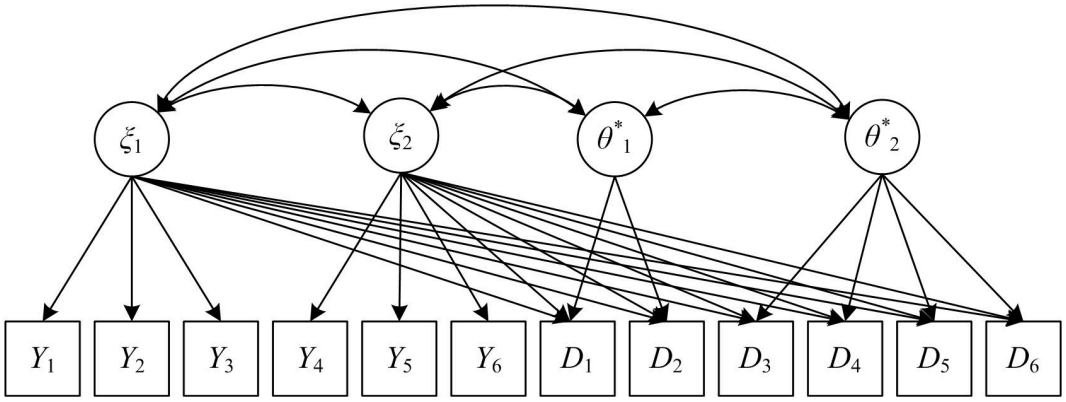

Figure 4.20: Graphical representation of the W$_{Dif}$MIRT Rasch model. All discrimination parameters represented by single-headed arrows are fixed to one.

applicable if the response indicators D_i in the equivalent BMIRT Rasch model indicate only a single latent dimension θ_l. The model specification becomes more difficult if the variables D_i are indicators of more than one latent response propensity θ_l. Such cases with a complex dimensionality will be examined below (see page 205).

The alternative Rasch-equivalent W$_{Res}$MIRT model for uni- and multidimensional variables ξ and θ Note that all latent dimensions in the 1PL-W$_{Dif}$MIRT model are allowed to be correlated. Otherwise, inappropriate restrictions are introduced in the model, since difference variables do typically correlate with the subtrahend and the minuend. Some authors proposed an alternative WMIRT model with the correlation $Cor(\xi, \tilde{\theta}) = 0$ (e. g. Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999). Indeed, such a model can also be derived. At first this will be done for the case of 1PL models. The resulting model is called the Rasch-equivalent W$_{Res}$MIRT model. The altered notation $\tilde{\theta}$, instead of θ or θ^*, indicates that the latent variable constructed in this model is different from that in the previous models. The restriction $Cor(\xi_m, \tilde{\theta}_l) = 0$ does not mean that the resulting model is more restrictive than the 1PL-BMIRT or the 1PL-W$_{Dif}$MIRT model. Rather $\tilde{\theta}$ is defined as variable, which is always regressively independent and therefore uncorrelated with all variables ξ_m (with $m = 1, \ldots, M$): The

residual of the regression $E(\theta \mid \boldsymbol{\xi})$[19]. The definition of latent variables as residuals is not new. A well-known application is to model method effects as latent residuals in confirmatory factor analysis (e. g. Geiser & Lockhart, 2012, February 6). A residual is only defined with respect to a particular regression. In the case of WMIRT models for nonignorable missing data, that is the regression of the latent response propensity on the latent ability. In the remainder, this model will be denoted as W_{Res}MIRT model. The concrete model specification is easy in the case when ξ and θ are each unidimensional, but might be less obvious in models where $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are multidimensional variables. The different model specifications will be derived next, starting with the Rasch-equivalent W_{Res}MIRT model for the case of undimensional latent variables ξ and θ.

A distinctive property of the BMIRT and W_{Dif}MIRT Rasch model examined previously is that all discrimination parameters α_{im} and γ_{il} are equal to one. It can be shown that this restriction is incompatible with the construction of $\tilde{\theta}$ as a residual. At least some of the discrimination parameters need to be freely estimable parameters, while the $Cor(\xi, \tilde{\theta})$ is fixed to be zero. For that reason, the model derived here is denoted as the Rasch-equivalent W_{Res}MIRT model. The general model equation of the logit vector $\boldsymbol{l(Y, D)}$ in this model is also given by Equation 4.80. If the measurement model of ξ based on $\boldsymbol{Y}$ without missing data is the Rasch model, then $\boldsymbol{\alpha}$ is the same in all three models - the 1PL-BMIRT model, the 1PL-W_{Dif}model, and the Rasch-equivalent W_{Res}MIRT model. Hence, all α_{im} are set to zero or one in advance. Note that the equality of $\boldsymbol{\alpha}$ in all equivalent models is a necessary but insufficient condition to ensure the equivalent construction of $\boldsymbol{\xi}$. The derivation of the Rasch-equivalent W_{Res}MIRT model reveals that the restriction $Cor(\xi, \tilde{\theta}) = 0$ requires that the elements $\tilde{\gamma}_{im}$ of $\tilde{\boldsymbol{\gamma}}_{\xi}$ of Equation 4.80 are estimable parameters. Note that the symbol $\sim$ denotes the parameters of the Rasch-equivalent W_{Res}MIRT model that differs from that if the alternative BMIRT- and W_{Dif}MIRT Rasch model. Let $\theta = E(\theta \mid \xi) + \zeta$, with the linear regression $E(\theta \mid \xi) = b_0 + b_1\xi$ and the residual $\zeta = \theta - E(\theta \mid \xi)$. Inserting the model equation of the latent response propensity into the logit equation of the manifest response indicators gives

$$
\begin{aligned}
l(D_i) &= \theta - \gamma_{i0} && (4.105)\\
&= E(\theta \mid \xi) + \zeta - \gamma_{i0} && (4.106)\\
&= b_0 + b_1\xi + \zeta - \gamma_{i0}. && (4.107)
\end{aligned}
$$

[19] In general a regression $E(Y \mid \boldsymbol{X})$ and the residual $\varepsilon = Y - E(Y \mid \boldsymbol{X})$ are always uncorrelated. For a proof see (Steyer & Eid, 2001; Steyer, 2002)

Defining $\tilde{\theta} = \zeta$ and setting $b_1 = \tilde{\gamma}_{i\xi}$ gives

$$l(D_i) = \tilde{\theta} + \gamma_{i\xi}\xi - (\gamma_{i0} - b_0), \tag{4.108}$$

with $\tilde{\gamma}_{i0} = \gamma_{i0} - b_0$ as the thresholds of the response indicator variables in the Rasch-equivalent W_{Res}MIRT model. This equation applies to all response indicators D_i Hence, the discrimination parameters $\gamma_{i\xi}$ are equal for all I response indicators. Hence, all parameters $\tilde{\gamma}_{i\xi}$, with $i = 1, \ldots, I$, need to be equal but freely estimable in application. That requires constrained parameter estimation with respect to the elements of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{\xi}}$. Hence, equality constraints needs to be specified in the model.

Before this alternative model is also applied to Data Example A, the Rasch-equivalent W_{Res}model will be generalized to multidimensional latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. In that case, the regression $E(\boldsymbol{\theta}\,|\,\boldsymbol{\xi})$ is multivariate. If $\boldsymbol{\theta}$ is P-dimensional, then the regression $E(\boldsymbol{\theta}\,|\,\boldsymbol{\xi})$ consists of P regressions $E(\theta_l\,|\,\boldsymbol{\xi})$, with $l = 1, \ldots, P$. If the latent dimensions θ_l and ξ_m are linearly regressively dependent, then the covariances $Cov(\theta_l, \xi_m)$ can alternatively be modeled by the P multiple linear regressions $E(\theta_l\,|\,\boldsymbol{\xi}) = b_0 + \sum_{m=1}^{M} b_{lm}\xi_m$. Hence, let $\theta_l = E(\theta_l\,|\,\xi_m) + \zeta_l$ with ζ_l the residual. Replacing θ_l in Equation 4.92 by the regression and its residual yields

$$l(D_i) = E(\theta_l\,|\,\xi_m) + \zeta_l - \gamma_{i0} \tag{4.109}$$

$$= b_0 + \sum_{m=1}^{M} b_{lm}\xi_m + \zeta_l - \gamma_{i0}. \tag{4.110}$$

Analogous to Equation 4.108, $\tilde{\theta}_l = \zeta_l$ and $b_{lm} = \tilde{\gamma}_{im}$. Hence,

$$l(D_i) = \sum_{m=1}^{M} \gamma_{im}\xi_m + \tilde{\theta}_l - \tilde{\gamma}_{i0}. \tag{4.111}$$

with the thresholds $\tilde{\gamma}_{i0} = b_0 - \gamma_{i0}$. Accordingly, the model equation for the response indicators is

$$P(D_i = 1\,|\,\tilde{\theta}_l, \xi) = \frac{\exp(\sum_{m=1}^{M} \tilde{\gamma}_{im}\xi_m + \tilde{\theta}_l - \tilde{\gamma}_{i0})}{1 + \exp(\sum_{m=1}^{M} \tilde{\gamma}_{im}\xi_m + \tilde{\theta}_l - \tilde{\gamma}_{i0})}. \tag{4.112}$$

This equation holds for all response indicators that indicate the latent dimension θ_l. Hence, all discrimination parameters $\tilde{\gamma}_{im}$ of the response indicators that constitute the measurement model of θ_l in the 1PL-BMIRT model are equal to b_{lm} in the Rasch-equivalent W_{Res}model. Therefore, the parameters $\tilde{\gamma}_{im}$ need to be constrained to be equal in appli-

cation. However, only those elements in $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{\xi}}$ that indicate the same latent dimension θ_l are equal. This is at least the case if there is a simple structure in the measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D}$ alone. Hence, if the response indicators D_i indicate more than one latent dimension θ_l in the BMIRT model, then the implied restrictions and equalities are more complex. Such cases with complex dimensionality will be considered below (see page 205).

For illustration, the equivalent W_{Res}MIRT Rasch model of the hypothetical model given by Figure 4.14 and 4.20 is displayed in Figure 4.21. In this example, two latent variables $\tilde{\theta}_1$ $\tilde{\theta}_2$ are required, which are defined as the residuals ζ_1 and ζ_1 of the two multiple regressions

$$E(\theta_1 \,|\, \xi_1, \xi_2) = \mathrm{b}_{10} + b_{11}\xi_1 + b_{12}\xi_1 \tag{4.113}$$

$$E(\theta_2 \,|\, \xi_1, \xi_2) = \mathrm{b}_{20} + b_{21}\xi_1 + b_{22}\xi_1 \tag{4.114}$$

Due to the equality $b_{lm} = \tilde{\gamma}_{im}$ for all response indicators constituting the measurement model of θ_l, the following equalities result for the hypothetical example displayed in Figure 4.21:

$$\begin{aligned} b_{11} &= \tilde{\gamma}_{11} = \tilde{\gamma}_{21} \\ b_{12} &= \tilde{\gamma}_{12} = \tilde{\gamma}_{22} \\ b_{21} &= \tilde{\gamma}_{31} = \tilde{\gamma}_{41} = \tilde{\gamma}_{51} = \tilde{\gamma}_{61} \\ b_{22} &= \tilde{\gamma}_{32} = \tilde{\gamma}_{42} = \tilde{\gamma}_{52} = \tilde{\gamma}_{62} \end{aligned} \tag{4.115}$$

In applications, these equalities need to be imposed by the use of equality constraints. In contrast, the sub-matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$ of $\boldsymbol{\Lambda}$ do not consist of estimable parameters. They must be priorly set to zero or one as in the BMIRT Rasch model. Additionally, all covariances $Cov(\xi_m, \tilde{\theta}_l)$ need to be fixed to zero. In contrast, there are no restrictions with respect to the covariances $Cov(\xi_m, \xi_w)$ $(m \neq w)$ and $Cov(\tilde{\theta}_l, \tilde{\theta}_k)$ $(l \neq k)$ that are freely estimable parameters.

Application of the Rasch-equivalent W_{Res}MIRT model to Data Example A The Rasch-equivalent W_{Res}MIRT model is Rasch-equivalent but strictly speaking not a multidimensional Rasch model since the discrimination parameters $\tilde{\gamma}_{im} \neq 1$ need to be estimated. Therefore, software for two-parameter models are required that allow for constrained parameter estimation. If the Rasch-equivalent W_{Res}MIRT is equivalent to the

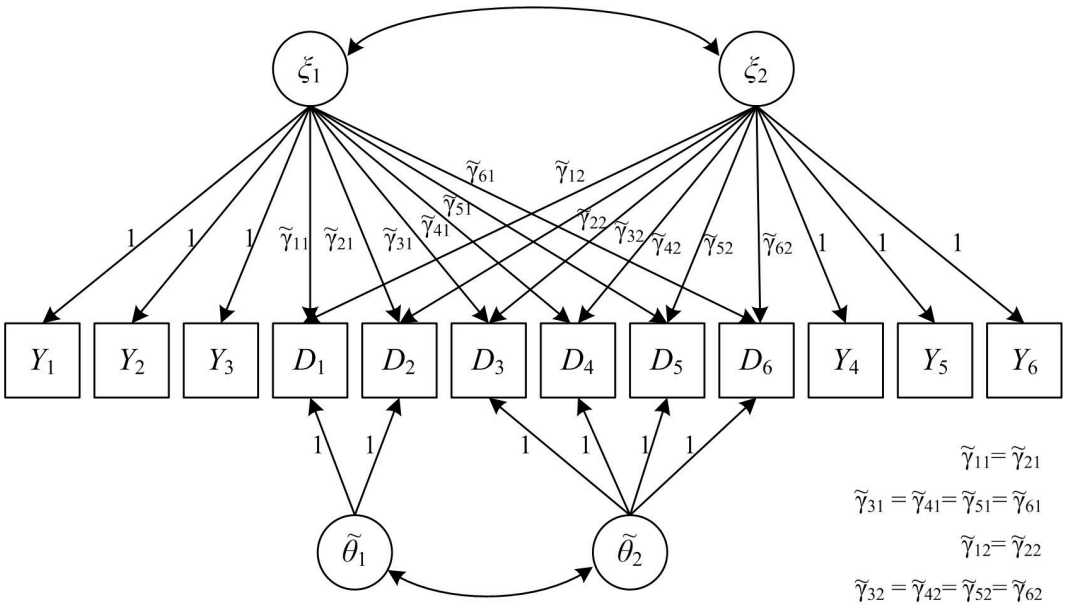

Figure 4.21: Graphical representation of the Rasch-equivalent W_{Res}MIRT model.

1PL-BMIRT and 1PL-W_{Dif}model, then the item and person parameter estimates in these models should be equal and the goodness-of-fit should be identical. M*plus* (Muthén & Muthén, 1998 - 2010) was used for parameter estimation. The input file for Data Example A is shown in Appendix B (see Listing A.3). Not all available software for MIRT allow for imposing equality constraints with respect to the item discriminations. In this situation, a relaxed version of the W_{Res}MIRT Rasch model can be applied alternatively. In this model, the equality constraints are left out. If the 1PL-BMIRT Rasch model is appropriate and the equality-constraints of the Rasch-equivalent W_{Res}MIRT model are not specified, then the estimates $\hat{\tilde{\gamma}}_{im}$ are freely estimated but should be close to the theoretically implied values, which are the regression coefficient b_{lm}. Hence, the relaxed Rasch-equivalent W_{Res}MIRT model is unnecessarily liberal. In turn, however, if the Rasch-equivalent W_{Res}MIRT Rasch model does not fit to the data, then substantial differences in the parameter estimates may result since the relaxed W_{Res}MIRT Rasch model is less restrictive. For the sake of comparison, the relaxed Rasch-equivalent MIRT model was also applied to Data Example A. The number of estimated parameters is higher in this model and has, therefore, less degrees of freedom. Accordingly, the relaxed W_{Res}MIRT Rasch model cannot be equivalent to the 1PL-BMIRT or the 1PL-W_{Dif}MIRT model in terms of model fit. Table 4.11 gives goodness of fit indices of the four different models applied to Data Example A: (1) the BMIRT Rasch model, (2) the W_{Dif}MIRT Rasch model, (3) the W_{Res}MIRT

Rasch model and (4) the relaxed W_{Res}MIRT Rasch model. Apparently, the deviation of the estimates of model implied and true response pattern probabilities are equal for the BMIRT Rasch model, the W_{Dif}MIRT Rasch model, and the Rasch-equivalent W_{Res}MIRT model. Since the relaxed Rasch-equivalent W_{Res}MIRT model is less restrictive, the log-likelihood is higher, indicating better model fit. However, the model is unnecessarily complex, indicated by higher information criteria, compared to the more restrictive but more parsimonious MIRT Rasch models and the Rasch-equivalent W_{Res}-MIRT model. The estimated item difficulties of 1PL-BMIRT and Rasch-equivalent W_{Res}MIRT were al-

Table 4.11: Goodness-of-fit Indices of the BMIRT-, W_{Dif}MIRT-, the Rasch-Equivalent W_{Res}MIRT-, and the Relaxed Rasch-equivalent W_{Res}MIRT Model (Data Example A).

Model	log-ℓ	n_{par}	AIC	BIC
BMIRT Rasch	-46535.705	63	93197.410	93550.267
W_{Dif}MIRT Rasch model	-46535.705	63	93197.410	93550.267
Rasch-eq. W_{Res}MIRT model	-46535.706	63	93197.411	93550.268
Relaxed Rasch-eq. W_{Res}MIRT model	-46519.046	92	93222.092	93737.375

Note: n_{par} = Number of estimated parameters.

most identical. Only one item (Y_3) showed a difference in the third decimal place. The estimates $\hat{\beta}_i$ obtained in the relaxed Rasch-equivalent W_{Res}MIRT model were also very close to those of the 1PL-BMIRT model. The absolute differences between the estimates of both models ranged between zero and 0.019 with the mean of 0.006. Hence, the estimates are practically the same. M*plus* was used for parameter estimation of the (relaxed) Rasch-equivalent W_{Res}MIRT model. This program allows only for EAP person parameter estimation. For that reason, the equivalence of the construction of ξ in these models are demonstrated using EAPs exclusively. In Figure 4.22, the EAPs obtained by the Rasch-equivalent W_{Res}MIRT model and the relaxed version of this model are compared with the EAPs estimated in the 1PL-BMIRT model. The correlation is close to one in both cases. The MSE of the EAPs was 0.222 in the Rasch-equivalent W_{Res}MIRT model and 0.223 in the relaxed Rasch-equivalent W_{Res}MIRT model. The mean of the absolute difference in the EAPs of the latent residual $\tilde{\theta}$ estimated in both models was 0.029. The correlation was $r = 0.999$. Hence, the estimates were practically identical as well.

Due to the results, it is concluded that the 1PL-BMIRT model, the 1PL-W_{Dif}MIRT model, and the Rasch-equivalent W_{Res}MIRT model are equivalent with respect to (a) the construction of ξ, (b) the adjustment of bias due to missing responses, and (c) the model fit. The relaxed Rasch-equivalent W_{Res}MIRT model is only equivalent with respect to (a)

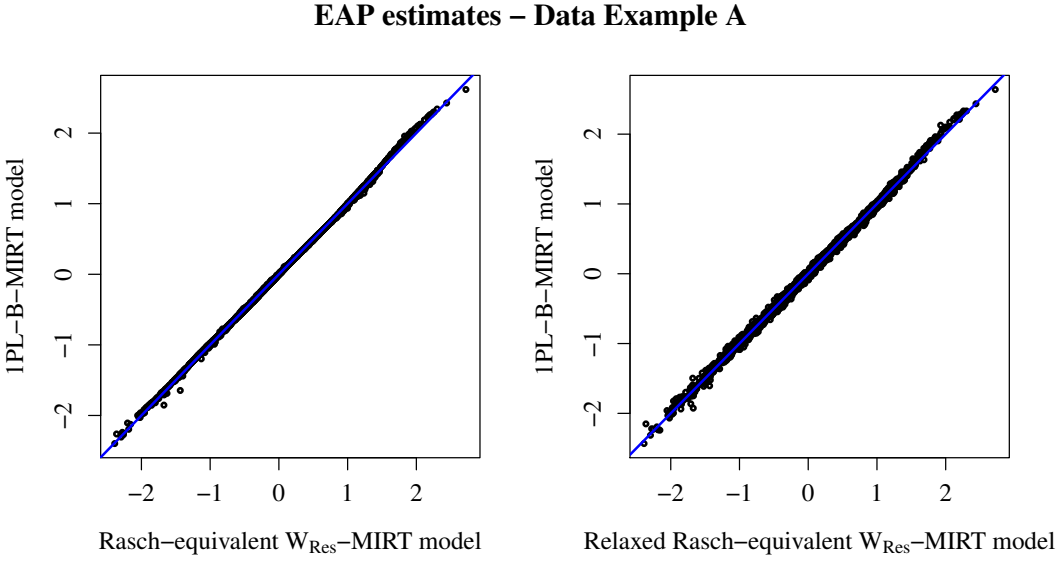

Figure 4.22: Comparison of EAP person parameter estimates of the BMIRT Rasch model, the Rasch-equivalent W_{Res}MIRT model (left) and the relaxed Rasch-equivalent W_{Res}MIRT (right). The blue line are the regression lines.

and (b), but not in terms of model fit.

Two-parameter MIRT models for nonignorable missing data with complex dimensionality In this section, the MIRT models for nonignorable missing data are generalized to (a) two-parameter models and (b) to cases with complex dimensional structure of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. The term complex dimensional structure refers to within-item multidimensionality of items Y_i in the measurement model of $\boldsymbol{\xi}$ and within-item multidimensionality of D_i in the measurement model of $\boldsymbol{\theta}$. Such a case is illustrated by the artificial example displayed in Figure 4.23. Compared with Figure 4.14, some of the response indicators D_i indicate more than one latent dimension θ_l. Similarly, there are test items Y_i indicating more than one latent ability ξ_m. Hence, within item-multidimensionality with respect to some manifest variables exists, even if the measurement models of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are considered separately. The general model equations of Y_i and D_i given by the Equations 4.77 and 4.78 remain valid in such cases. Note that the abbreviation BMIRT model does not mean that the items Y_i and D_i are between-item dimensional. This term refers to the conditional stochastic independencies $\boldsymbol{Y} \perp \boldsymbol{\theta}|\boldsymbol{\xi}$ and $\boldsymbol{D} \perp \boldsymbol{\xi}|\boldsymbol{\theta}$ reflected by the structure of $\boldsymbol{\Lambda}$ with $\boldsymbol{\gamma}_\xi = \mathbf{0}$ in BMIRT models (see Equations 4.80 and 4.82). Hence, the matrix $\boldsymbol{\Lambda}$ of discrimination parameters is block-diagonal. Only the sub-matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_\theta$ consist of

estimable parameters. The interpretation of the latent variables θ_l is essentially the same as in the BMIRT model with a simple structure. Given that $\gamma_{il} \geq 0$, higher values of θ_l means higher probabilities to respond to those items i whose response indicators D_i indicate θ_l given the other dimensions $\theta_{h \neq l}$. However, the parameters γ_{il} are nothing else than partial logistic or probit regression coefficients, and it is generally possible that some parameters $\gamma_{il} < 0$, indicating that the probability of an item response decreases when θ_l increases *given* the other dimensions $\theta_{h \neq l}$. Despite such peculiarities, $\boldsymbol{\theta}$ is interpreted as a multidimensional latent response propensity variable in the 2PL-BMIRT model. In

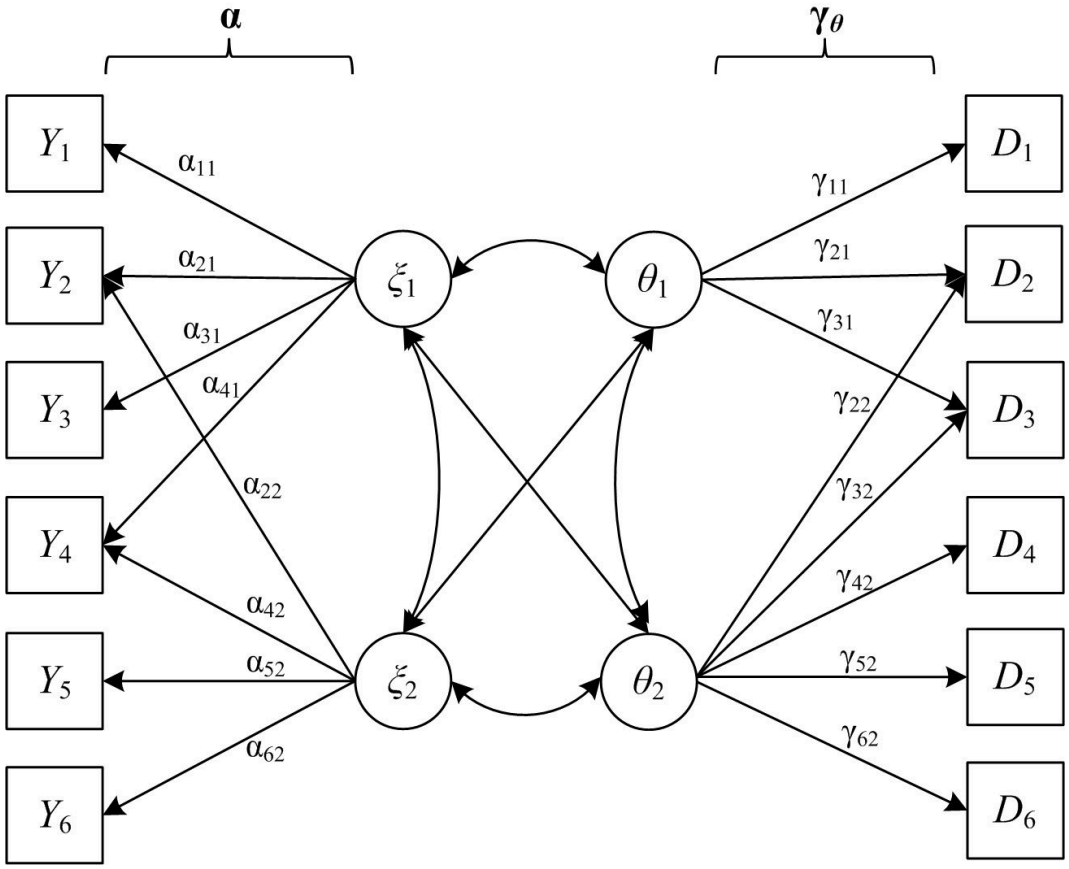

Figure 4.23: MIRT model with within-item multidimensional items Y_i and response indicators D_i (2PL-BMIRT model).

contrast to the BMIRT Rasch model, not all elements α_{im} and γ_{il} are fixed to zero or one prior to the analysis. Only some of these parameters are fixed to zero if the respective item Y_i or response indicator D_i does not indicate the latent dimension ξ_m or θ_l directly. The 2PL models need additional restrictions for model identification. At least one of the discrimination parameters α_{im} is fixed to a particular value, or the variance $Var(\xi_m)$ is

fixed. Accordingly, at least one γ_{il} is fixed, or the variance $Var(\theta_l)$ is fixed to a value greater than one. As in the 1PL-models, the location of the latent variables is identified by fixing at least one threshold per dimension or assigning a arbitrary value to $E(\xi_m)$ and $E(\theta_l)$. Hence, the application and specification of 2PL-BMIRT models in cases with complex dimensionality is straightforward and does not require further clarification. This is quite different for equivalent 2PL-WMIRT models that are derived next. In order to demonstrate the application of 2PL-BMIRT and 2PL-WMIRT models in M*plus* a further simulated data example, which is called Data Example C, was used. Data Example C is described in detail in Appendix 3. M*plus* input files as well as summaries of essential results are presented in Appendix 3. The specification of the 2PL-BMIRT model of Figure 4.23 is shown in Listing A.9.

As in the case of one-parameter MIRT models for nonignorable missing data, equivalent 2PL-WMIRT models are rationally derived, starting from the 2PL-BMIRT model. In a first step, this will be done specifically for the hypothetical model displayed in Figure 4.23. Afterwards, general specification rules will be derived for equivalent 2PL-W$_{Dif}$MIRT and 2PL-W$_{Res}$MIRT models.

Derivation of the 2PL-W$_{Res}$MIRT model considering complex dimensionality As in the Rasch-equivalent W$_{Res}$MIRT model, the p-dimensional latent variable $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2)$ is defined as the multivariate residual $\boldsymbol{\zeta} = \zeta_1, \ldots, \zeta_P$, with $\zeta_l = \theta_l - E(\theta_l \,|\, \boldsymbol{\xi})$. In the hypothetical example given in Figure 4.23, the two regressions

$$E(\theta_1 \,|\, \xi_1, \xi_2) = b_{10} + b_{11}\xi_1 + b_{12}\xi_2 \tag{4.116}$$

and

$$E(\theta_2 \,|\, \xi_1, \xi_2) = b_{20} + b_{21}\xi_1 + b_{22}\xi_2 \tag{4.117}$$

are involved. Thus, in a joint bivariate regression, the two-dimensional residual is $\boldsymbol{\zeta} = (\zeta_1, \zeta_2)$. An alternative 2PL-W$_{Res}$MIRT model can be derived setting $\tilde{\boldsymbol{\theta}} = \boldsymbol{\zeta}$ with $\tilde{\theta}_l = \zeta_l$. If the dimensional structure is complex since manifest variables Y_i and D_i indicator more than one latent dimension ξ_m or θ_l respectively, then the logit equations of $l(Y_i)$ and $l(D_i)$

in the 2PL-BMIRT model are

$$l(Y_i) = \sum_{m=1}^{M} \alpha_{im}\xi_m - \beta_i \tag{4.118}$$

$$l(D_i) = \sum_{l=1}^{P} \gamma_{il}\theta_l - \gamma_{i0}. \tag{4.119}$$

Hence, the logits are linear combinations of the respective latent dimensions. The model equations of $l(Y_i)$ are the same as in the target model which is the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$. In order to derive the equivalent 2PL-W$_{Res}$MIRT model, the latent response propensity dimensions θ_l in Equation 4.119 are replaced by the respective regressions $E(\theta_l \,|\, \boldsymbol{\xi}) + \zeta_l$. In the further derivations it is assumed that all θ_l are linear in $\xi_1, \ldots, \xi_M$[20]. In that case, the joint distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ of the latent variables can be modelled by P multiple linear regression $E(\theta_l \,|\, \boldsymbol{\xi})$ and the respective residuals ζ_l, with $l = 1, \ldots, P$. Returning to the example of Figure 4.23, there are response indicators that are between-item multidimensional as D_1 and two response indicators, D_2 and D_3, that are within-item multidimensional. For the further derivations, the first and the second response indicators are used exemplarily. According to Equation 4.119, the logit equations of these two variables are

$$l(D_1) = \gamma_{11}\theta_1 - \gamma_{10} \tag{4.120}$$

$$l(D_2) = \gamma_{21}\theta_1 + \gamma_{22}\theta_2 - \gamma_{20}. \tag{4.121}$$

The latent propensity dimensions θ_1 and θ_2 can be replaced by their constituting parts - the regressions given in Equations 4.116 and 4.116 and the corresponding residuals ζ_1 and ζ_2, yielding

$$\begin{aligned} l(D_1) &= \gamma_{11}[E(\theta_1 \,|\, \xi_1, \xi_2) + \zeta_1] - \gamma_{10} \\ &= \gamma_{11}[b_{10} + b_{11}\xi_1 + b_{12}\xi_1 + \zeta_1] - \gamma_{10} \\ &= \gamma_{11}b_{11}\xi_1 + \gamma_{11}b_{12}\xi_2 + \gamma_{11}\zeta_1 - (\gamma_{10} - \gamma_{11}b_{10}), \end{aligned} \tag{4.122}$$

[20] Unfortunately, currently available software packages do not allow for non-linear regressions between latent variables in MIRT models.

and

$$
\begin{aligned}
l(D_2) &= \gamma_{21}[E(\theta_1 \,|\, \xi_1, \xi_2) + \zeta_1] + \gamma_{22}[E(\theta_2 \,|\, \xi_1, \xi_2) + \zeta_2] - \gamma_{20} \qquad (4.123)\\
&= \gamma_{21}[b_{10} + b_{11}\xi_1 + b_{12}\xi_2 + \zeta_1] + \gamma_{22}[b_{20} + b_{21}\xi_1 + b_{22}\xi_2 + \zeta_2] - \gamma_{20}\\
&= \gamma_{21}b_{11}\xi_1 + \gamma_{21}b_{12}\xi_2 + \gamma_{21}\zeta_1 + \gamma_{22}b_{21}\xi_1 + \gamma_{22}b_{22}\xi_2 + \gamma_{22}\zeta_2\\
&\quad -(\gamma_{20} - \gamma_{22}b_{20} - \gamma_{21}b_{10}).
\end{aligned}
$$

To obtain an applicable model, the equations need to be rearranged so that each latent variable appears once. This condition is already met for the first response indicator since D_1 is between-item multidimensional in the 2PL-BMIRT model. However, rearranging the terms of logit equation of the second response indicator gives

$$
\begin{aligned}
l(D_2) &= (\gamma_{21}b_{11} + \gamma_{22}b_{21})\xi_1 + (\gamma_{21}b_{12} + \gamma_{22}b_{22})\xi_2 + \gamma_{21}\zeta_1 + \gamma_{22}\zeta_2 \qquad (4.124)\\
&\quad -(\gamma_{20} - \gamma_{22}b_{20} - \gamma_{21}b_{10}).
\end{aligned}
$$

To obtain the final 2PL-W$_{Res}$MIRT model, the variables ζ_l are replaced by $\tilde{\theta}_l$. The final logit equations of the model are

$$
\begin{aligned}
l(D_1) &= \underbrace{\gamma_{11}b_{11}}_{\tilde{\gamma}_{11}}\xi_1 + \underbrace{\gamma_{11}b_{12}}_{\tilde{\gamma}_{21}}\xi_2 + \gamma_{11}\tilde{\theta}_1 - \underbrace{(\gamma_{10} - \gamma_{11}b_{10})}_{\tilde{\gamma}_{10}} \qquad (4.125)\\
l(D_2) &= \underbrace{(\gamma_{21}b_{11} + \gamma_{22}b_{21})}_{\tilde{\gamma}_{12}}\xi_1 + \underbrace{(\gamma_{21}b_{12} + \gamma_{22}b_{22})}_{\tilde{\gamma}_{22}}\xi_2 + \gamma_{21}\tilde{\theta}_1 + \gamma_{22}\tilde{\theta}_2 \qquad (4.126)\\
&\quad - \underbrace{(\gamma_{20} - \gamma_{22}b_{20} - \gamma_{21}b_{10})}_{\tilde{\gamma}_{20}}.
\end{aligned}
$$

It can be seen that the parameters $\tilde{\gamma}_{im}$ of $\tilde{\boldsymbol{\gamma}}$ are functions of both the parameters γ_{il} of $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$ and the regressions coefficients b_{lm}. Furthermore, the thresholds $\tilde{\gamma}_{i0}$ are functions of γ_{i0}, γ_{il}, and the intercepts b_{l0}. These functional dependencies between the model parameters need to be taken into account in application of the 2PL-W$_{Res}$MIRT model by specification of non-linear constraints. The increased complexity due to within-item multidimensionality becomes obvious when comparing Equations 4.125 and 4.126. With increasing numbers of latent dimensions θ_l that are indicated by a response indicator D_i, and increasing latent dimensions ξ_m, the more complex the implied non-linear constraints with respect to $\tilde{\gamma}_{im}$ and $\tilde{\gamma}_{i0}$ are. The general form of these non-linear constraints will be derived next for the case of a P-dimensional latent variable $\tilde{\boldsymbol{\theta}}$ and an M-dimensional latent ability $\boldsymbol{\xi}$. In this case, there are P multiple linear regressions $E(\theta_l \,|\, \boldsymbol{\xi}) = b_{l0} + \sum_{m=1}^{M} b_{lm}\xi_m$ and the P-

dimensional residual $\boldsymbol{\zeta} = \zeta_1, \ldots, \zeta_P$. Inserting θ_l in Equation 4.119 by the $E(\theta_l \,|\, \boldsymbol{\xi}) + \zeta_l$ gives

$$\begin{aligned} l(D_i) &= \sum_{l=1}^{P} \gamma_{il} \Big[E(\theta_l \,|\, \boldsymbol{\xi}) + \zeta_l \Big] - \gamma_{i0} \\ &= \sum_{l=1}^{P} \gamma_{il} \Big[b_{l0} + \sum_{m=1}^{M} b_{lm} \xi_m \Big] + \sum_{l=1}^{P} \gamma_{il} \zeta_l - \gamma_{i0} \\ &= \sum_{l=1}^{P} \sum_{m=1}^{M} \gamma_{il} b_{lm} \xi_m + \sum_{l=1}^{P} \gamma_{il} \zeta_l + \sum_{l=1}^{P} \gamma_{il} b_{l0} - \gamma_{i0}. \end{aligned} \tag{4.127}$$

Finally, the notation is adapted. In the 2PL-W$_{Res}$MIRT model, $\zeta_l = \tilde{\theta}_l$. The general model equation of the logits of the response indicators D_i is then

$$l(D_i) = \sum_{m=1}^{M} \tilde{\gamma}_{im} \xi_m + \sum_{l=1}^{P} \gamma_{il} \tilde{\theta}_l - \tilde{\gamma}_{i0}, \tag{4.128}$$

with the discrimination parameters

$$\tilde{\gamma}_{im} = \sum_{l=1}^{P} \gamma_{il} b_{lm}, \tag{4.129}$$

and the thresholds

$$\tilde{\gamma}_{i0} = \gamma_{i0} - \sum_{l=1}^{P} \gamma_{il} b_{l0}. \tag{4.130}$$

The parameter estimation in the 2PL-W$_{Res}$MIRT model requires for constrained optimization with respect to the parameters in the likelihood $\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d}; \boldsymbol{\iota}, \boldsymbol{\phi})$. The non-linear constraints with respect to elements of $\boldsymbol{\gamma_\xi}$ of Λ (see Equation 4.80) and the thresholds $\tilde{\gamma}_{i0}$ are given by the Equations 4.129 and 4.130. In application, it is recommended to identify the model in a way that simplifies the model-implied constraints. For example, if the expected values are fixed to $E(\xi_m) = E(\theta_l) = 0$, then all intercepts b_{l0} are zero as well. It follows that $\tilde{\gamma}_{i0} = \gamma_{i0}$. In this case, only the constraints with respect to the item discrimination parameters, that involve the partial regression coefficients, need to be specified. Software, such as M*plus*, allows to define the coefficients b_{lm} implicitly as new parameters. This is illustrated by the M*plus* input file (see Listing A.10) in Appendix 5.3, which refers to the fictional model depicted in Figure 4.24. The regressions $E(\theta_l \,|\, \boldsymbol{\xi})$ are not depicted in

the graph. Note that all covariances $Cov(\xi_m, \tilde{\theta}_l)$ are equal to zero. This restriction must also be made explicitly in M*plus*. If the equality constraints with respect to elements of

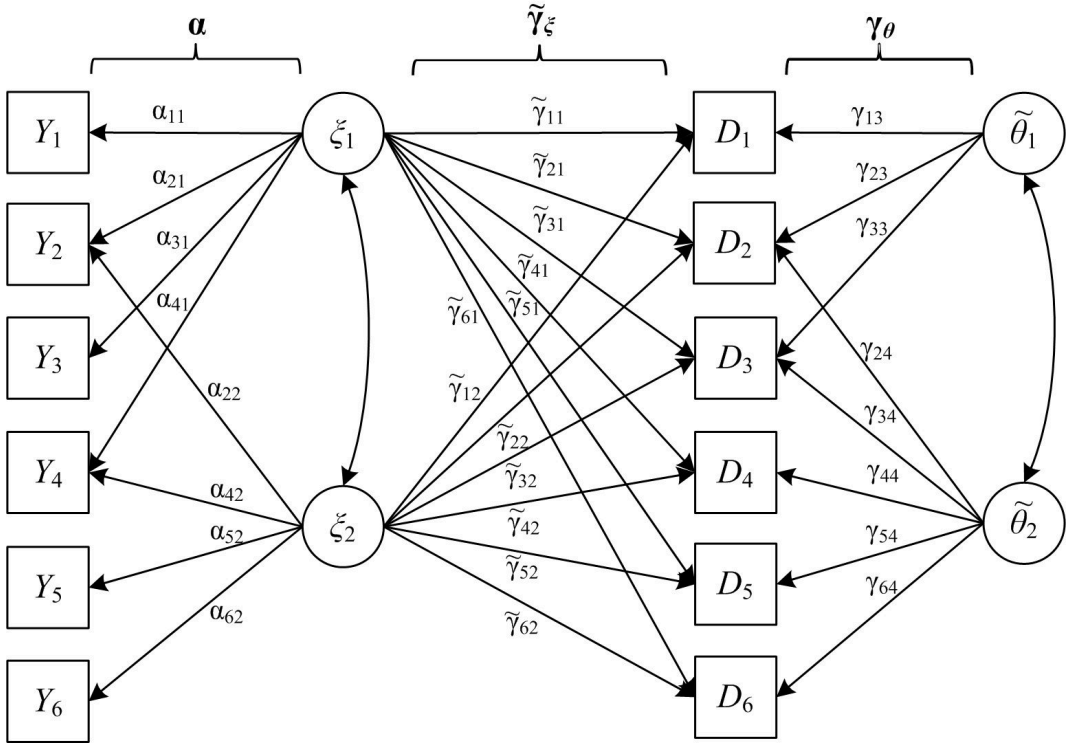

Figure 4.24: Graphical representation of the 2PL-W$_{Res}$MIRT model.

$\tilde{\boldsymbol{\gamma}}$ are not specified, so that all parameters $\tilde{\gamma}_{im}$ are freely estimable parameters, the *relaxed* 2PL-W$_{Res}$MIRT model results. This model is applicable in software packages that do not allow for the specification of non-linear constraints. As in the case of the relaxed Rasch-equivalent W$_{Res}$MIRT model, the relaxed 2PL-W$_{Res}$MIRT model is less restrictive and, therefore, not equivalent to the 2PL-BMIRT and the 2PL-W$_{Dif}$MIRT model in terms of model fit. However, if the latter fit the data well the item and person parameter estimates of all MIRT models including the relaxed 2PL-W$_{Res}$MIRT model should be close.

Derivation of the 2PL-W$_{Dif}$MIRT model considering complex dimensionality A generalized two-parameter model can be derived with a latent difference variable $\boldsymbol{\theta}^*$ instead of a latent response propensity $\boldsymbol{\theta}$ or the latent residual $\tilde{\boldsymbol{\theta}}$. In the W$_{Dif}$Rasch model, the idea has been developed to define $\boldsymbol{\theta}^* = \theta_1^*, \ldots, \theta_P^*$ as a multidimensional latent difference variable with $\theta_l^* = \theta_l - \sum_{m=1}^{M} \xi_m$ for all $l = 1, \ldots, P$. Following this approach, a 2PL-W$_{Dif}$MIRT model can be derived which is equivalent to the 2PL-BMIRT model.

As the 2PL-W$_{Res}$MIRT model, the general 2PL-W$_{Dif}$MIRT model allows for complex dimensionality due to within-item multidimensionality in response indicators D_i. Again, the model can be derived from the 2PL-BMIRT model. The model equation of the logits $l(D_i)$ is given in Equation 4.119. In the 2PL-W$_{Dif}$MIRT model, the latent response propensity dimensions θ_l are replaced by the difference variables θ_l^* as defined above, so that

$$l(D_i) = \sum_{m=1}^{M} \gamma_{im}^* \xi_m + \sum_{l=1}^{P} \gamma_{il} \theta_l^* - \gamma_{i0}^*. \tag{4.131}$$

Since each dimension θ_l^* is defined as the difference $\theta_l - \sum_{m=1}^{M} \xi_m$, this expression can be inserted into Equation 4.131, yielding

$$l(D_i) = \sum_{m=1}^{M} \gamma_{im}^* \xi_m + \sum_{l=1}^{P} \gamma_{il} \left[\theta_l - \sum_{m=1}^{M} \xi_m \right] - \gamma_{i0} \tag{4.132}$$

$$= \sum_{m=1}^{M} \gamma_{im}^* \xi_m + \sum_{l=1}^{P} \gamma_{il} \theta_l - \sum_{l=1}^{P} \sum_{m=1}^{M} \gamma_{il} \xi_m - \gamma_{i0}. \tag{4.133}$$

In this equation, the logit $l(D_i)$ is a function of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. Since the resulting 2PL-W$_{Dif}$MIRT model should be equivalent to the 2PL-BMIRT model, the conditional stochastic independence $D_i \perp \boldsymbol{\xi} \,|\, \boldsymbol{\theta}$ needs to be true. That means that $l(D_i)$ is only a function of $\boldsymbol{\theta}$ but not of $\boldsymbol{\xi}$. Furthermore, the equality of Equation 4.133 with Equation 4.119 needs to be preserved. This is the case if γ_{im}^* is set equal to the sum $\sum_{l=1}^{P} \gamma_{il}$. Inserting this expression in Equation 4.133 for all γ_{im}^* gives

$$l(D_i) = \sum_{m=1}^{M} \sum_{l=1}^{P} \gamma_{il} \xi_m + \sum_{l=1}^{P} \gamma_{il} \theta_l - \sum_{l=1}^{P} \sum_{m=1}^{M} \gamma_{il} \xi_m - \gamma_{i0} \tag{4.134}$$

$$= \sum_{l=1}^{P} \gamma_{il} \theta_l - \gamma_{i0}. \tag{4.135}$$

This is exactly the equation of the logit $l(D_i)$ in the 2PL-BMIRT model. Conclusively, the 2PL-W$_{Dif}$MIRT model is obtained if all parameters γ_{im}^* of $\boldsymbol{\gamma}_{\boldsymbol{\xi}}$ are constraint, so that

$$\gamma_{im}^* = \sum_{l=1}^{P} \gamma_{il}. \tag{4.136}$$

The thresholds γ_{i0} are the same in both models and can be freely estimated respectively. Similarly to the 2PL-W$_{Res}$MIRT model, the functional relation between the model parameters γ^*_{im} and γ_{il} needs to be imposed in the model for parameter estimation in real applications. This requires the specification of linear constraints with respect to γ^*_{im}, which are given by Equation 4.136. In Appendix 5.3, the M*plus* input file of the 2PL-W$_{Dif}$MIRT model is provided (see Listing A.11), which is graphically represented in Figure 4.25. This model refers to the hypothetical example, which is also illustrated in Figures 4.23 and 4.24.

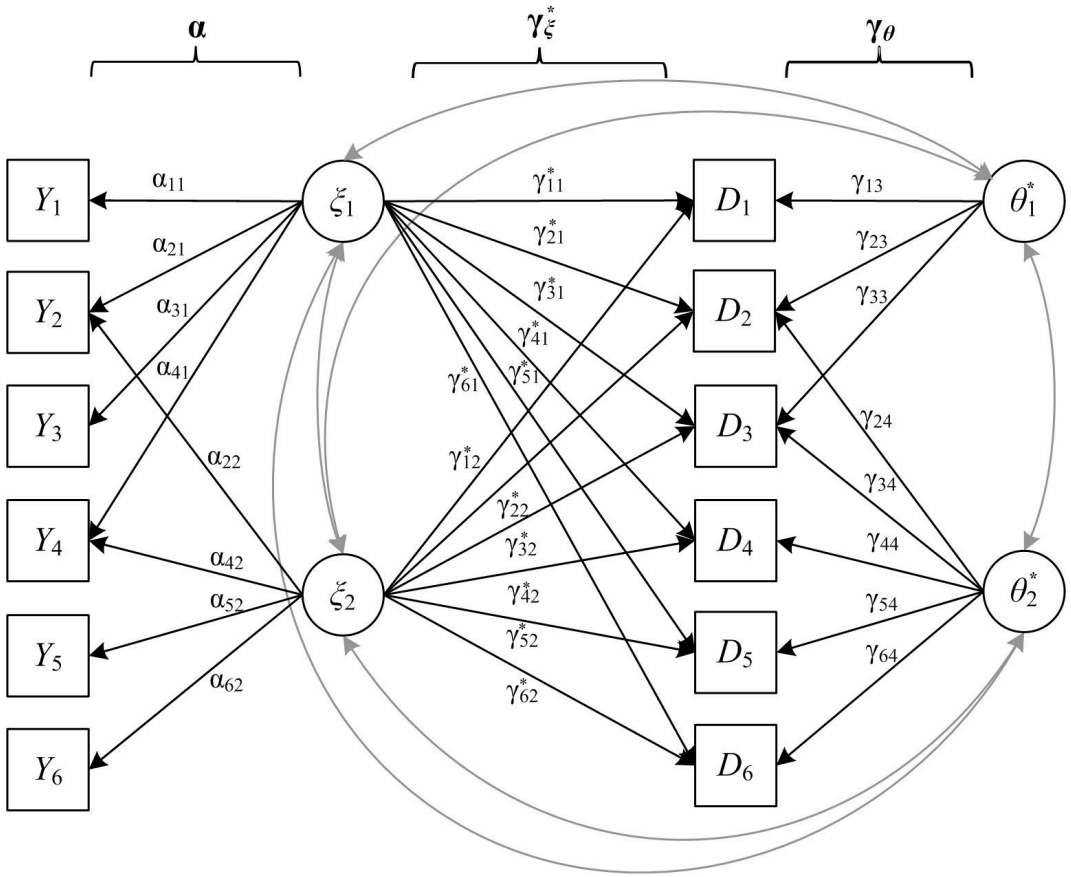

Figure 4.25: Graphical representation of the 2PL-W$_{Dif}$MIRT model. The covariances are represented by grey double-headed arrows.

Interim conclusion Here a brief summary of this section is given. Different MIRT models that account for nonignorable missing data have been developed. For the case of the Rasch-MIRT model for nonignorable missing data, three 1PL-WMIRT models have been derived that are equivalent to the the 1Pl-BMIRT model. These are (a) the 1PL-W$_{Dif}$MIRT

model, (b) the Rasch-equivalent W_{Res}MIRT model, and (c) the relaxed Rasch-equivalent W_{Res}MIRT model. The models (a) and (b) are equivalent with respect to three criteria which have been turned out to be essential in missing data models. These are (1) equivalent construction of $\boldsymbol{\xi}$, (2) equivalent adjustment for missing data, and (3) equal model-fit . The relaxed Rasch-equivalent W_{Res}MIRT model is only equivalent with respect to the criteria (1) and (2) but not in terms of model fit, given the 1PL-BMIRT model and 1PL-W_{Dif}model fit the data. In this case, the relaxed Rasch-equivalent is unnecessarily liberal since more parameters than required are freely estimated. In a subsequent step, the 2PL-MIRT models for missing data were developed starting form the 2PL-BMIRT model. It was emphasized that the term between-item dimensional MIRT models is used here in a slightly different way. It refers to the conditional stochastic independence of the response indicators D_i of the latent ability $\boldsymbol{\xi}$ given the latent response propensity $\boldsymbol{\theta}$. This is the distinctive feature of the BMIRT models compared to the derived WMIRT models. Finally, three 2PL-WMIRT models were derived: (a) the 2PL-W_{Dif}MIRT model, (b) the 2PL-W_{Res}MIRT model, and (c) the relaxed 2PL-W_{Res}MIRT model. The models (a) and (b) are equivalent to the 2PL-BMIRT model with respect to the three criteria of model equivalence in IRT models for missing data. The relaxed 2PL-W_{Res}MIRT model is only equivalent with respect to the construction of $\boldsymbol{\xi}$ and the adjustment for missing data but not in terms of model fit. This is analogous to the relaxed Rasch-equivalent W_{Res}MIRT model. In general, the WMIRT models differ from the BMIRT models in the construction of the latent variables. Whereas only in the BMIRT models a potentially multidimensional latent response propensity is constructed, a latent difference variable or a latent residual is constructed in the W_{Dif}MIRT and the W_{Res}MIRT models respectively. To allow for complex dimensional structures underlying $\boldsymbol{Y}$ and $\boldsymbol{D}$, the difference variable $\boldsymbol{\theta^*}$ in the W_{Dif}MIRT models was defined as the multidimensional difference variable $\theta_1^*, \ldots, \theta_P^*$ with $\theta_l^* = \theta_l - \sum_{m=1}^{M} \xi_m$. The latent residual $\tilde{\boldsymbol{\theta}}$ in the W_{Res}MIRT models was defined as the residual $\boldsymbol{\zeta} = \boldsymbol{\theta} - E(\boldsymbol{\theta} \,|\, \boldsymbol{\xi})$. Hence, $\tilde{\theta}_l = \zeta_l = \theta_l - E(\theta_l \,|\, \boldsymbol{\xi})$. The question may arise which of these models should be preferred in application. One of the major purposes of this work was to show under which circumstances the models are equivalent. If they are equivalent, then it does not matter which model is chosen. In real applications, the available software may limit the model choice. To the best knowledge of the author, only M*plus* allows to fit all the strongly equivalent models introduced here. The BMIRT models do not require non-linear constraints and are, therefore, applicable in most MIRT software. For example, the 1PL-BMIRT model can also be applied using ConQuest. Additionally, BMIRT models are easy to specify and the parameters are easy to interpret. Therefore,

this class of models is be recommended for most applications. However, MIRT software for bi-factor analysis allows only for uncorrelated latent variables. In such cases, at least the W_{Res}MIRT models are applicable. If the specification of non-linear constraints are impossible, then the resulting models refer to the relaxed Rasch-equivalent WMIRT or the relaxed 2PL-W_{Res}MIRT model. Note, that the results and parameter estimates of the BMIRT models and the respective relaxed Rasch-equivalent WMIRT or relaxed 2PL-W_{Res}MIRT model will only substantially differ if the models are misspecified.

Here it is argued that the BMIRT models should be preferred for three reasons. First, these models are the easiest to specify, since they do not require for linear or non-linear constraints. Second, only in the BMIRT models can the latent variables θ_l of $\boldsymbol{\theta}$ be interpreted as latent response propensities. The correlations between the latent ability dimensions ξ_m and θ_l might be of additional diagnostic value. The same information is also contained in the WMIRT models but much harder to extract especially in models with a complex multidimensionality. Third, Hooker, Finkelman, and Schwartzman (2009) demonstrated, mathematically and empirically, that statistical correct estimation procedures of MIRT models without a simple structure can produce paradoxical results especially in the person parameter estimates obtained by within-item multidimensional models. These findings were confirmed by Hooker (2010) and Finkelman, Hooker, and Wang (2010). To avoid such results, which are nearly impossible to detect in real applications, the BMIRT models should be preferred. In all models discussed here, it is essential that the dimensionality of $\boldsymbol{\theta}$ is appropriately taken into account. In almost all publictions of MIRT models for nonignorable missing data, unidimensionality of the latent response propensity is assumed. However, here it is argued that an ignored multidimensionality can mask the nonignorability of the missing data mechanism and can impede adjustment for missing responses. This will be demonstrated in the subsequent section.

In conclusion, because of the simplicity, clearness, and its ease of use, the BMIRT Rasch model and its generalization, the 2PL-BMIRT model, are recommended in real applications. But even these models require knowledge about the measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D}$ which should be carefully examined, especially with respect to the underlying dimensionality. This will be examined next.

4.5.3.4 Dimensionality of the Latent Response Propensity

In the previous derivations of the MIRT models for nonignorable missing responses, it was is emphasized that the latent response propensity $\boldsymbol{\theta}$ and, therefore, the latent difference variable $\boldsymbol{\theta}^*$ and the latent residual $\tilde{\boldsymbol{\theta}}$ are potentially multidimensional. Why is

that? In most published applications the latent response propensity was assumed to be unidimensional (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). In none of these publications did the dimensionality underlying $\boldsymbol{D}$ seem to be critically challenged. However, here it is argued that the appropriate model of $\boldsymbol{D}$ is essential to correct for nonignorable missing data in $\boldsymbol{Y}$. In this section, it will be demonstrated that a joint measurement model of $(\boldsymbol{Y}, \boldsymbol{D})$ fails to correct for nonignorable missing data, if unidimensionality of θ is falsely assumed although a multidimensional latent variable $\boldsymbol{\theta}$ underlies $\boldsymbol{D}$. For that reason, a further data example was simulated, denoted as Data Example B in the remainder. $I = 30$ manifest dichotomous items Y_i were simulated with a single latent variable ξ. The Rasch model was chosen for all items Y_i with the same item difficulties as in Data Example A (see Table 3.1). The latent response propensity $\boldsymbol{\theta}$ was chosen to be two-dimensional in Data Example B. The response indicators $D_1 - D_{20}$ constitute the measurement model of θ_1, and $D_{21} - D_{30}$ indicate θ_2. Thus, the measurement model of $\boldsymbol{\theta}$ follows a simple structure. The thresholds γ_{i0} of the response indicators were the same as in Data Example A (see Table 3.1). All item discriminations γ_{i1} were fixed to one. As in Data Example A, item nonresponses are generally more likely for more difficult items than for easier items. The joint distribution $g(\xi, \boldsymbol{\theta})$ of the three latent variables was specified as a multivariate normal distribution, with

$$\begin{pmatrix} \xi \\ \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0.8 \\ 0 & 1 & 0 \\ 0.8 & 0 & 1 \end{pmatrix}\right]. \tag{4.137}$$

Hence, only the second latent dimension θ_l was correlated with the latent ability ξ. Due to stochastic independence of all manifest variables given the latent variables, it follows: $\xi \perp \theta_1 \Rightarrow Y_i \perp D_i$ for $i = 1, \ldots, 20$. This implies that the missing data mechanism with respect to the first 20 items is MCAR. In contrast, the items $Y_{21} - Y_{30}$ suffer from a nonignorable missing data mechanism, since $\xi \not\perp \theta_2 \Rightarrow Y_i \not\perp D_i$. Hence, only the probability of missing responses in the last 10 items is stochastically dependent on the latent ability ξ. However, this implies a nonignorable missing data mechanism w.r.t. $\boldsymbol{Y}$ since conditional stochastic independence $\boldsymbol{D} \not\perp \boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs}$ results. The sample size was $N = 1000$.

The overall proportion of missing data in Data Example B was 46.97%. The response rates per item ranged between 19.50 - 99.00% ($M = 53.03\%$, $SD = 21.34\%$). The correlation between the proportion correct P^+ and the proportion of answered items $\bar{D}$ was $r = 0.193$ ($t = 6.216$, $df = 998$, $p < 0.001$). In real application, this would be strong ev-

idence for non-ignorability of the missing data mechanism. However, the model-implied correlation between P^+ and $\bar{D}$ of only the first 20 items is zero. In Data Example B, the sample correlation is even slightly negative $r = -0.122$ (t = -3.877, $df = 998$, $p < 0.001$) due to sampling error . According to the positive correlation $cor(\xi, \theta_2) = 0.8$, the sample correlation between P^+ and $\bar{D}$ of the last 10 items is $r = 0.329$ ($t = 10.679$, $df = 941$, $p < 0.001$).

In order to study the effect of ignoring the dimensionality of $\boldsymbol{\theta}$, four models were applied. At first, the unidimensional complete data model based on $\boldsymbol{Y} = \boldsymbol{y}$ for reasons of comparison. The same model was applied to the incomplete data $\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}$. In this model, the nonignorability of the missing data mechanism was not taken into account. The third model was the between-item multidimensional IRT model based on $(\boldsymbol{Y}_{obs}, \boldsymbol{D}) = (\boldsymbol{y}_{obs}, \boldsymbol{d})$, with only a single latent variable θ. Hence, the measurement model of the latent response propensity was misspecified in this model. Finally, the correct measurement model with three latent variables ξ, θ_1 and θ_2 was fitted to the data $(\boldsymbol{y}_{obs}, \boldsymbol{d})$. The 2PLM was chosen for data analysis to study the impact of the model choice with respect to the discrimination parameters. In none of the four models was the mean bias of the estimates $\hat{\alpha}_i$ significantly different from 1[21]. The mean squared error of $\hat{\alpha}_i$ was the lowest in the complete data model ($MSE = 0.009$). In both, the unidimensional 2PLM ignoring missing data and the 2PL-BMIRT models, the MSE was 0.040. In the correct specified measurement model with a two-dimensional latent response propensity, the MSE was 0.033. The results are in line with the findings presented in Section 3.2.3; The item discrimination estimates are not systematically biased even if the missing data mechanism was NMAR. Therefore, the estimates $\hat{\alpha}_i$ will not be further considered here. In contrast, the item difficulties were found to be biasedly estimated in presence of nonignroable missing responses (see Section 3.2.2). Hence, in Data Example B, the estimates $\hat{\beta}_i$ were expected to be underestimated as well. The upper triangle of the matrix plot in Figure 4.26 compares the true and the estimated item difficulties of the four models applied to Data Example B. The unfilled circles refer to the items $Y_1 - Y_{20}$ with ignorable missing data. The filled triangles indicate the true and estimated item difficulties of the items $Y_{21} - Y_{30}$ with nonignorable missing data. In general, the bias was small in all four models. However, the expected systematic underestimation in the last ten items with nonignorable missing data could be confirmed when the missing data where ignored. The mean bias $Bias(\hat{\beta})$ was

[21] Complete data model: $Bias(\hat{\alpha}) = -0.002$ ($t = -0.144, df = 29, p = 0.887$); 2PLM ignoring missing data: $Bias(\hat{\alpha}) = 0.013$ ($t = 0.346, df = 29, p = 0.732$); Misspecified 2PL-BMIRT: $Bias(\hat{\alpha}) = 0.013$ ($t = 0.360, df = 29, p = 0.721$; Correctly specified 2PL-BMIRT: $Bias(\hat{\alpha}) = 0.002$ ($t = 0.076, df = 29, p = 0.940$)

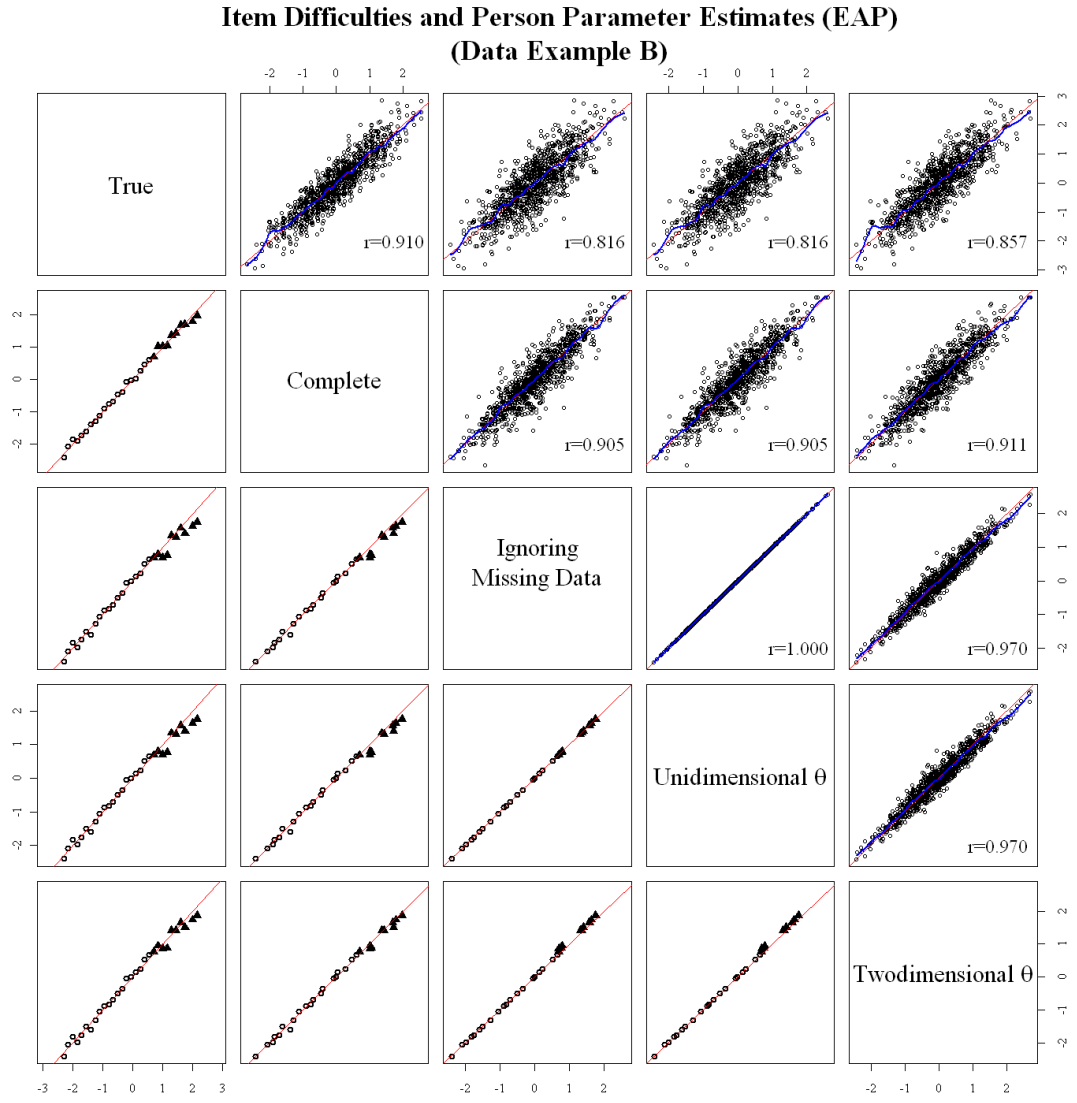

Figure 4.26: Comparison of the true values of ξ and β_i for Data Example B with corresponding estimates obtained by different models. The red lines represent the bisectric. The blue lines are smoothing spline regressions.

significantly different from zero in (a) the unidimensional model ignoring missing data ($Bias(\hat{\boldsymbol{\beta}}) = 0.061$ ($t = 2.103, df = 29, p = 0.044$), and (b) the misspecified 2PL-BMIRT model ($Bias(\hat{\boldsymbol{\beta}}) = 0.060$ ($t = 2.049, df = 29, p = 0.049$). In contrast the mean bias in the correctly specified 2PL-BMIRT was $Bias(\hat{\boldsymbol{\beta}}) = 0.023$ ($t = 0.963, df = 29, p = 0.344$) and the *MSE* has been reduced to 0.028. For reasons of comparison, the mean bias of the complete data model was $Bias(\hat{\boldsymbol{\beta}}) = 0.002$ ($t = 0.107, df = 29, p = 0.915$) and the *MSE* = 0.007.

The most remarkable finding is that the item and EAP person parameter estimates of the undimensional model that ignores missing data and the misspecified 2PL-BMIRT model are practically identical (see upper triangle of the matrix plot in Figure 4.26). The joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$ seems not to have any effect on parameter estimation. Why is that? A closer look revealed that the the discrimination estimates $\hat{\gamma}_{21;\theta} - \hat{\gamma}_{30;\theta}$ in the misspecified 2PL-BMIRT model ranged only between -0.081 and 0.156. The mean -0.034 was not significantly different from zero ($t = -1.337$, $df = 9$, $p = 0.214$) and none of the single estimates $\hat{\gamma}_{21;\theta} - \hat{\gamma}_{30;\theta}$ were significantly different from zero. In contrast, the estimates $\hat{\gamma}_{1;\theta} - \hat{\gamma}_{20;\theta}$ ranged between 0.958 and 1.340, with the mean 1.071 ($t = 50.926$, $df = 19$, $p < 0.001$). This is close to the true value 1 that was used for data simulation. The results imply that the single latent variables θ in the misspecified undimensional 2PL-BMIRT model is almost exclusively constructed based on the response indicators D_1 to D_{20}. As a consequence θ mostly represents θ_1 and not θ_2. Accordingly, the estimated correlation between ξ and θ in the misspecified 2PL-BMIRT model was $r = 0.020$ ($SE = 0.062$, $t = 0.325$,$p = 0.745$). If ξ and θ are independent, then parameter estimation hardly benefits from the joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$. This is obvious considering EAP person parameter estimation. The prior $g(\xi, \theta)$ used in the EAP estimation (see Equation 4.90) can be written as $g(\xi \,|\, \theta)g(\theta)$. Given $\xi \perp \theta$, it follows that $g(\xi \,|\, \theta)g(\theta) = g(\xi)g(\theta)$. The distribution of ξ is equal for each value of θ. Hence, $\boldsymbol{D}$ and therefore θ do not contain any additional information with respect to ξ given $\boldsymbol{Y}$. In other words, the misspecified 2PL-BMIRT model in this example works as though the missing data would be MCAR. In fact, in real application an applied researcher could be tempted to conclude that the missing data mechanism is ignorable, since the estimated correlation between θ and ξ was not significantly different from zero. If the model would be correctly specified, then that implies stochastic independence between $\boldsymbol{Y}$ and $\boldsymbol{D}$.

In the simulated Data Example B, especially the EAP estimates profited from the correct specification of the measurement model of $\boldsymbol{\theta}$. The estimated correlations in the correctly specified 2PL-BMIRT were $r(\xi, \theta_1) = 0.036$ ($SE = 0.042$, $t = 0.855$,$p = 0.392$),

$r(\xi, \theta_2) = 0.738$ ($SE = 0.031$, $t = 24.148$,$p < 0.001$), and $r(\theta_1, \theta_2) - 0.038$ ($SE = 0.043$, $t = -0.875$,$p = 0.382$). Accordingly, the prior used to estimate persons' EAPs $\hat{\xi}$ in this particular example is $g(\xi, \theta_1, \theta_2)$. Since, $\xi \perp \theta_1$ and $\theta_1 \perp \theta_2$ it follows

$$g(\xi, \theta_1, \theta_2) = g(\xi \,|\, \theta_1, \theta_2) g(\theta_1, \theta_2) \tag{4.138}$$

$$= g(\xi \,|\, \theta_2) g(\theta_2) g(\theta_1) \tag{4.139}$$

It can be seen that the conditional distribution of ξ differs depending on θ_2. Hence, the EAP estimates $\hat{\xi}$ shrink approximately toward the conditional expected values $E(\xi \,|\, \theta_2 = \theta_2)$ in the correct 2PL-BMIRT model instead of toward $E(\xi)$. θ_2 is indicated by $D_{21} - D_{30}$. That is why $\boldsymbol{D}$ is informative with respect to the latent variable ξ. Exploiting this information leads to the correlation $r(\xi, \hat{\xi}) = 0.857$ that is larger than $r(\xi, \hat{\xi}) = 0.816$ in the misspecified model or the simple 2PLM that ignores missing data. It should be noted that the missing induced bias is comparably small in Data Example B. This is because of the 20 items in the test with missing responses that are MCAR. Observed responses to these items provide a lot of valuable information for item and person parameter estimation and limit the negative effects of nonignorable missing responses in the last ten items.

Based on Data Example B, it could be shown that the inclusion of $\boldsymbol{D}$ in a joint model of $(\boldsymbol{Y}, \boldsymbol{D})$ needs to be done appropriately. Disregarding the correct dimensionality of $\boldsymbol{\theta}$ will potentially lead to an MIRT model that can fail to correct the bias due to nonignorable missing data although $\boldsymbol{D}$ is included in the model. In application, the correct model for $\boldsymbol{D}$ needs to be found. Here it is argued that this task should involve all sources of information including explorative procedures to determine the number of dimensions θ_l. The reason is that the response indicators D_i are not items of a rationally constructed test. The number of dimensions underlying $\boldsymbol{D}$ and their substantial meaning can hardly be anticipated prior to application. In this respect, variables D_i differ from items Y_i that are constructed theoretically driven. Of course, practical experiences in applied testings and theoretical considerations may help to develop ideas about the dimensionality of the latent response propensity. For example, Rose et al. (2010) found that item characteristics can be related to the willingness to complete test items in PISA 2006. Whereas the mean response rates per item and the item means were correlated in open constructed-response items, this relation was negligible in multiple choice items, which show generally high response rates. This does not necessarily imply multidimensionality. However, if the willingness to respond to different item types varies across persons depending on the response format, then the resulting item-by-person interaction (2009) implies multidimensionality of

$\boldsymbol{\theta}$. The dimensionality of ξ might also provide information about the dimensionality of $\boldsymbol{\theta}$. Especially if the dimensions ξ_m and $\xi_{k \neq m}$ are only weakly correlated but the probability to omit items is strongly correlated with the respective latent ability, then the dimensionality of $\boldsymbol{\theta}$ might mimic the dimensionality of $\boldsymbol{\xi}$. In any case, the dimensionality of the latent response propensity should be checked and a suited model with respect to $\boldsymbol{\theta}$ needs to be specified. In Data Example B, for instance, a explorative factor analysis (EFA) for dichotomous variables (Jöreskog & Moustaki, 2000; Mislevy, 1986; Muthén, 1978) revealed the wrong dimensionality. The EFA provides eigenvalues, χ^2-values, root mean squared error of approximation (RMSEA), and further fit statistics that help to determine the number of required dimensions underlying $\boldsymbol{D}$. The difference between the EFA for continuous versus dichotomous manifest variables lies in the correlation matrix used for model estimation. Instead of Pearson correlations, the matrix of tetrachoric correlations is used in the case of dichotomous variables. Furthermore, suited least square estimators are used to estimate model parameters (Muthén, 1978, 1998 - 2004; Muthén, do Toit, & Spisic, 1997; Wirth & Edwards, 2007).

The EFA for dichotomous variables was applied to the data matrix $\boldsymbol{D} = \boldsymbol{d}$ of Data Example B. M*plus* 6 (Muthén & Muthén, 1998 - 2010) was used for model estimation. The mean and variance adjusted weighted least square (WLSMV) estimator was applied. Three models were fitted with one to three latent dimensions θ_l. Promax was used as the rotation method in order to allow for correlated factors. Table 4.12 gives the goodness of fit indices and Figure 4.27 shows the scree plot with the eigenvalues of the estimated tetrachoric correlation matrix of the 30 response indicators D_i. Note that the EFA with

Table 4.12: Model fit statistics for EFAs of the tetrachoric correlation matrix of response indicators (Data Example B).

Factors	χ^2	*df*	*p*-value	RMSEA	RMSR
1	1394.283	405	< 0.001	0.049	0.107
2	390.434	376	0.293	0.006	0.062
3	336.801	348	0.657	0.000	0.054

Note: RMSEA = Root mean squared error of approximation; RMSR = Root mean squared residual.

only one latent dimension is equivalent to a unidimensional CFA model identified by $E(\theta) = 0$ and $Var(\theta) = 1$. In line with the data generating models used in Data Example B, the two-dimensional model shows a considerable better model fit than the unidimensional model. The scree plot supports a solution with two factors. Additionally, in the two-factor model, the matrix of factor loadings approximately follows a simple structure with

the exception of D_1, which loads on both latent dimensions ($\lambda_{11} = 0.557$, $\lambda_{21} = 0.468$). The loadings of the response indicators $D_2 - D_{20}$ pertaining to the first latent dimension ranged between $0.415 - 0.575$, and between $-0.132 - 0.053$ with respect to the second latent dimension. Conversely, the loadings of the variables $D_{21} - D_{30}$ were located between $-0.070 - 0.089$ with respect to the first dimension, and $0.410 - 0.556$ with regard to the second dimension. In the three factor solution only the single variable D_1 had a substantial factor loading on the third dimension $\lambda_{31} = 1.077$. The pattern of loadings with respect to the first two factors was preserved. This is in line with the existing literature. As Reckase (2009) found, the factor structure of a model with too many dimensions embeds the dimensional structure with the required number of dimensions. However, EFA for

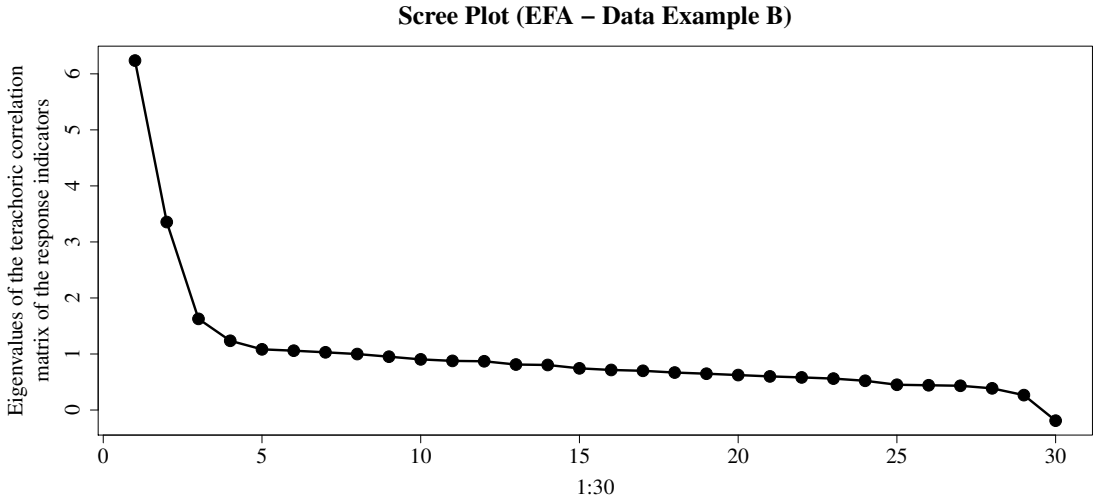

Figure 4.27: Screeplot based on the tetrachoric correlation matrix of response indicators (Data Example B).

categorical vaiables is only one approach to study the dimensionality in IRT measurement models. Many other methods for the empirical assessment of the underlying dimensional structure of a test consisting of dichotomously or ordered categorical scored items have been developed (Jasper, 2010; Reckase, 2009; Roussos, Stout, & Marden, 1998; Stout et al., 1996; Tate, 2003). It is far beyond the scope of this work to review these methods here. The major focus of this section was to illustrate the importance of the correct specification of the measurement model of $\boldsymbol{\theta}$. However, increased dimensionality in joint measurement models of $(\boldsymbol{Y}, \boldsymbol{D})$ can become numerically challenging. The development of simpler but sufficient model-based approaches for nonignorable missing data would be of great value. The latent regression model and a multiple group model for nonignorable missing data can

be alternatives to MIRT models in some applications. In these models, the measurement model of $\boldsymbol{\theta}$ can be omitted. Nevertheless, both approaches - the latent regression models and the multiple group IRT models - require knowledge about the required number of dimensions that sufficiently explain the stochastic dependencies between the response indicators D_i.

4.5.4 Latent Regression IRT Models for Nonignorable Missing data

The major disadvantage of different between- and the within-item multidimensional IRT models is their complexity. The number of manifest variables is doubled due to the inclusion of the response indicators D_i in a joint measurement model of $(\boldsymbol{Y}, \boldsymbol{D})$. If additionally the underlying dimensional structure of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ is complex and the number of latent dimensions ξ_m and θ_l is high, then the analysis becomes computationally demanding and very time consuming. As Cai (2010) stated, high dimensional MIRT models are still computationally challenging. Less complex models for nonignorable missing responses might be preferable in such situations. Using the PISA 2006 data, Rose, von Davier, and Xu (2010) could show that substantially simpler models can reduce the bias due to nonignorable missing data equally well. They proposed a latent regression model (LRM) and a multiple group (MG) IRT model for item nonresponses that are NMAR. Both approaches are justified and examined here in more detail. The relation of these methods to the MIRT models introduced above will be outlined. Furthermore, the LRM and MG-IRT models for missing responses are also conceptually close to IRT models for missing responses that are MAR given a covariate $\boldsymbol{Z}$ (see Section 4.5.2). The basic idea is to use functions $f(\boldsymbol{D})$ of the response indicator vector as covariates in an LRM or as grouping variable in an MG-IRT model. Although $\boldsymbol{D}$ is taken into account to adjust for nonignorable missing data, the measurement model is considerably slimmed down to the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$. Therefore, the LRM and the MG-IRT models are much less complex compared to the MIRT models described previously. The two approaches are developed step by step starting with the LRM.

The general LRM for nonignorable missing data In the LRM proposed by Rose et al. (2010), the proportion of completed items $\bar{D}$ was used as predictor in a latent regression $E(\xi \mid \bar{D})$, with $\bar{D} = I^{-1} \sum_{i=1}^{I} D_i$, which was computed for each test taker. Formally, $\bar{D}$ is a function $f(\boldsymbol{D})$ of the response indicator vector $\boldsymbol{D}$. Other functions such as the sum score $\sum_{i=1}^{I} D_i$ might also be suited. The choice of the function $f(\boldsymbol{D})$ depends on many factors. For example, if there are several sub-tests that refer to different domains, then

the use of a single proportion of completed items can be improper. Instead, the proportion of responded or omitted items can be determined for each sub-test. In this case the latent regression model for item nonresponses becomes a multiple linear regression with the functions $f_j(\boldsymbol{D})$ of $\boldsymbol{f}(\boldsymbol{D}) = f_1(\boldsymbol{D}), \ldots, f_J(\boldsymbol{D})$ as regressors. For the case of a multidimensional latent variable $\boldsymbol{\xi}$, the most general form of the structural model of the latent regression approach for missing responses proposed here is

$$E[\boldsymbol{\xi} \,|\, \boldsymbol{f}(\boldsymbol{D})]. \tag{4.140}$$

It is important that the parameters $\boldsymbol{\iota}$ of the measurement model of $\boldsymbol{\xi}$ are jointly estimated with the parameters of the latent regression model.

Relation to MIRT models for nonignorable missing data The latent regression model and the MIRT models for nonignorable missing data are theoretically closely related. This will be demonstrated for the case of the between-item multidimensional MIRT model. From the basic model assumption of the MIRT models for nonignorable models (see Equations 4.5.4 - 4.74) follows, in general, $\boldsymbol{D} \perp \boldsymbol{Y} \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$, and in B-MIRT models $\boldsymbol{D} \perp \boldsymbol{Y} \,|\, \boldsymbol{\theta}$ holds. Assuming that the latent response propensity $\boldsymbol{\theta}$ would be a manifest variable included in the model as an auxiliary variable, the missing data mechanism w.r.t. $\boldsymbol{Y}$ would be MAR given $\boldsymbol{\theta}$. In this case, $\boldsymbol{\theta}$ would be a covariate like other variables represented by $\boldsymbol{Z}$. In Section 4.5.2, it was shown that the LRMs can be used to account for missing data that are MAR given $\boldsymbol{Z}$. Accordingly, an LRM with $E(\boldsymbol{\xi} \,|\, (\theta))$ would sufficiently account for item nonresponses if the B-MIRT model assumptions hold true. Furthermore, if (θ) were observable, then it were not required to be measured by $\boldsymbol{D}$. Hence, the response indicator vector $\boldsymbol{D}$ could be ignored. This can also be shown more formally considering ML estimation. Assuming that $\boldsymbol{\theta}$ is given, the full likelihood $\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi})$ is proportional to the joint distribution of $(\boldsymbol{D}, \boldsymbol{Y}, \boldsymbol{\theta})$, so that

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) \quad \propto \quad g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}, \boldsymbol{\phi}). \tag{4.141}$$

Using the factorization $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ (see Section 4.5.1) that is

$$\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) \quad \propto \quad g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}, \boldsymbol{\phi}). \tag{4.142}$$

As shown in Section 4.5.1, the likelihood of the observed data results from integrating over the distribution of $\boldsymbol{Y}_{mis}$. In this case, that is,

$$\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \int g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) d\boldsymbol{Y}_{mis}, \quad (4.143)$$

which can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto \int g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\theta} = \theta; \boldsymbol{\phi}) g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}) \int \Big\{ g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\theta} = \theta; \boldsymbol{\phi}) \\ &\quad g(\boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}) \Big\} d\boldsymbol{Y}_{mis} \end{aligned} \quad (4.144)$$

If the MIRT model assumptions hold true, from conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y} \,|\, \boldsymbol{\theta}$ implied by Equations 4.5.4 - 4.74 and local stochastic independence $Y_i \perp Y_j \,|\, \boldsymbol{\xi}$ holds, it follows that the observed data likelihood can be simplified to

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}) \int g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{\theta} = \theta; \boldsymbol{\phi}) g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{\theta} = \theta; \boldsymbol{\phi}) \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \end{aligned} \quad (4.145)$$

In this case, the integral is $\int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} = 1$, implying that the observed data likelihood is proportional to the product of the joint distribution of the observed partition of $\boldsymbol{Y}_{obs}$ and $\boldsymbol{\theta}$ and the conditional distribution of $\boldsymbol{D}$ given $\boldsymbol{\theta}$. This is

$$\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, \theta; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{\theta} = \theta; \boldsymbol{\phi}). \quad (4.146)$$

Hence, the likelihood can be factorized into two independent pieces with different sets of model parameters $\boldsymbol{\iota}$ and $\boldsymbol{\phi}$. Given the parameter spaces $\Omega_{\boldsymbol{\iota}}$ and $\Omega_{\boldsymbol{\phi}}$ are distinct, so that $\Omega_{\boldsymbol{\iota},\boldsymbol{\phi}} = \Omega_{\boldsymbol{\iota}} \times \Omega_{\boldsymbol{\phi}}$ the ignorability conditions hold in a joint model of $(\boldsymbol{Y}, \boldsymbol{\theta})$. Therefore, in application it is sufficient to maximize the observed data likelihood $\mathcal{L}(\boldsymbol{y}_{obs}, \theta; \boldsymbol{\iota})$, which is proportional to $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{\theta} = \theta; \boldsymbol{\iota})$, in order to obtain unbiased parameter estimates. Thus, if $\boldsymbol{\theta}$ would be available the inclusion of the complete response indicator $\boldsymbol{D}$ superfluous. Unfortunately, $\boldsymbol{\theta}$ is not observable in real applications. Nevertheless, a practical solution is to replace $\boldsymbol{\theta}$ by fallible measures of the true latent response propensity. Such proxies of $\boldsymbol{\theta}$ can be used as independent variables in the LRM for nonignorable missing

data. However, in a strict sense, it is assumed that $\boldsymbol{D}$ is conditionally stochastically independent of $\boldsymbol{Y}$ given the respective proxy of $\boldsymbol{\theta}$. Of course, it is possible that this assumptions does not hold even if $\boldsymbol{D} \perp \boldsymbol{Y} \,|\, \boldsymbol{\theta}$. However, if an appropriate proxy of $\boldsymbol{\theta}$ can be found, the remaining conditional stochastic dependency between $\boldsymbol{D}$ and $\boldsymbol{Y}$ given this proxy becomes negligible. Real data analyses have shown that the LRM for nonignorable missing data yields almost identical results compared to the MIRT models for nonignorable missing data. Furthermore, in the simulated Data Example C (see Appendix C) highly unreliable EAP estimates of two latent response propensities θ_1 and θ_2 have been used in an LRM. The resulting item and person parameter estimates turned out to be almost identical to the estimates of the 2PL-BMIRT model (see Figure 5.1).

Choosing functions $f(\boldsymbol{D})$ in the LRM for nonignorable missing data There are several candidates which could serve as proxies of $\boldsymbol{\theta}$. If θ is unidimensional latent variable constructed in a 1PLM or 2PLM based on $\boldsymbol{D}$, then the sum score $S_D = \sum_{i=1}^{I} D_i$ or the mean $\bar{D} = I^{-1} \sum_{i=1}^{I} D_i$ can simply be used (Rose et al., 2010). The larger the number of items is, the higher correlation between θ and S_D or $\bar{D}$ is, due to the increased reliability. $\bar{D}$ and S_D are simply the manifest test scores indicating the tendency to complete the items of the test. Thus, they serve as fallible measures of θ transformed into a different metric[22]. However, in the current work it was emphasized that $\boldsymbol{\theta}$ can be multidimensional. In such cases, the use of S_D or $\bar{D}$ might be an inappropriate oversimplification. To justify the suitability of the regressions $E(\boldsymbol{\xi} \,|\, S_D)$ or $E(\boldsymbol{\xi} \,|\, \bar{D})$, one needs knowledge of the dimensional structure underlying $\boldsymbol{D}$. Therefore, a stepwise procedure is recommended. First, the dimensionality of $\boldsymbol{\theta}$ is analyzed. Second, the appropriate functions $f(\boldsymbol{D})$ are chosen due to the result of the model for $\boldsymbol{D}$. If undimensionality holds true, then S_D or $\bar{D}$ can be used in the LRM. However, S_D or $\bar{D}$ can also be replaced by person parameter estimates $\hat{\theta}$ in the LRM since $\hat{\theta} = f(\boldsymbol{D})$. This is the recommended choice if $\boldsymbol{\theta}$ is multidimensional with a complex dimensional structure. If $\boldsymbol{\theta}$ is p-dimensional, then the estimate $\hat{\boldsymbol{\theta}} = \hat{\theta}_1, \ldots, \hat{\theta}_p$ is used in a multiple latent regression $E(\boldsymbol{\xi} \,|\, \hat{\boldsymbol{\theta}})$. If all response indicators are between-item multidimensional, so that each D_i is indicator of only a single dimension θ_l, then P sum scores S_{Dl} may be a viable alternative, where S_{Dl} is the sum of those response indicators D_i that are indicators of θ_l. It should also be mentioned that the regression $E(\boldsymbol{\xi} \,|\, \boldsymbol{D})$ is a special case of Equation 4.140. Hence, the latent ability can be regressed on all response indicators. In this case, the dimensionality of $\boldsymbol{\theta}$ needs not be studied. However, if the

[22] $\sum_{i=1}^{I} D_i = \sum_{i=1}^{I} P(D_i = 1 \,|\, \theta) + \sum_{i=1}^{I} \varepsilon_{D_i}$, with $\sum_{i=1}^{I} P(D_i = 1 \,|\, \theta)$ as the expected number of completed items which is a function $f(\theta)$. ε_{D_i} is the residual of the regression $P(D_i = 1 \,|\, \theta)$. Equivalently, $\bar{D} = I^{-1} \sum_{i=1}^{I} P(D_i = 1 \,|\, \theta) + I^{-1} \sum_{i=1}^{I} \varepsilon_{D_i}$, with $I^{-1} \sum_{i=1}^{I} P(D_i = 1 \,|\, \theta) = f(\theta)$.

number of variables becomes large, then the number of regression coefficients inflates and the model tends to be unnecessarily complex.

Note that the LRM allows for nonlinear regressions. Hence, the model can easily be extended to cases including polynomials of the functions $f(\boldsymbol{D})$. In this respect, the LRM is superior to the MIRT models discussed previously.

ML estimation of the LRM for nonignorable missing data So far, ML estimation in the LRM was only considered in order to demonstrate the relation to MIRT models for nonignorable missing data and the IRT models for missing data that are MAR given $\boldsymbol{Z}$. Now, ML estimation in the LRM for nonignorable missing data is generally considered. The derivations are quite close to the case where $\boldsymbol{\theta}$ was assumed to be known. Instead of the latent response propensity, a function $f(\boldsymbol{D})$ is considered. Accordingly, ML estimation rest upon the joint distribution of $(\boldsymbol{Y}, \boldsymbol{D}, f(\boldsymbol{D}))$. Note that conditional stochastic independence $\boldsymbol{Y} \perp f(\boldsymbol{D}) \,|\, \boldsymbol{D}$ always holds true, whereas the assumption $\boldsymbol{Y} \perp \boldsymbol{D} \,|\, f(\boldsymbol{D})$ needs not necessarily be true. The general complete data likelihood is

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}, \boldsymbol{d}, f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D} = \boldsymbol{d}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) \qquad (4.147) \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}). \end{aligned}$$

The observed data likelihood is again proportional to the integral over the missing variable $\boldsymbol{Y}_{mis}$. That is,

$$\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) \propto \int g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{D} = \boldsymbol{d}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) d\boldsymbol{Y}_{mis}. \quad (4.148)$$

The joint distribution can be factored yielding

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto \int \Big\{ g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\phi}) \qquad (4.149) \\ &\qquad g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \Big\} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}) \int \Big\{ g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\phi}) \\ &\qquad g(\boldsymbol{Y}_{mis} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}) \Big\} d\boldsymbol{Y}_{mis} \qquad (4.150) \end{aligned}$$

Further, it is assumed that local stochastic independence $Y_i \perp Y_j \,|\, \boldsymbol{\xi}$ holds true for all $i \neq j$. Additionally, if conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, (\boldsymbol{Y}_{obs}, f(\boldsymbol{D}))$ can be

assumed, then the observed data likelihood can be simplified to

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}_{obs}, \boldsymbol{d}, f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}) \\ &\quad \cdot \int g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\phi}) g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}) g(\boldsymbol{D} = \boldsymbol{d} \,|\, \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\phi}) \\ &\quad \cdot \int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis}.\end{aligned} \tag{4.151}$$

In this case, the last factor is $\int g(\boldsymbol{Y}_{mis}; \boldsymbol{\iota}) d\boldsymbol{Y}_{mis} = 1$ and does not affect ML parameter estimation. The likelihood can be factorized into two independent parts that can be maximized independently to yield unbiased parameter estimates $\hat{\boldsymbol{\iota}}$ and $\hat{\boldsymbol{\phi}}$, respectively. Hence, unbiased item and person parameters can be obtained by maximizing the *reduced* observed data likelihood

$$\mathcal{L}(\boldsymbol{y}_{obs}, f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) \propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}), \tag{4.152}$$

which merely includes the function $f(\boldsymbol{D})$ instead of $\boldsymbol{D}$. The most important characteristic is that the model of $\boldsymbol{D}$ represented by the parameter vector $\boldsymbol{\phi}$ is not involved anymore, which simplifies the model considerably. Since $f(\boldsymbol{D})$ is included as an exogenous variable in a regression, it is sufficient to model the conditional distribution $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs} \,|\, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota})$ instead of the joint distribution $g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota})$. If the test takers answered independently and local stochastic independence holds true, then the general MML function is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}_{obs}, f(\boldsymbol{d}); \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs} \,|\, f(\boldsymbol{D}) = f(\boldsymbol{d}); \boldsymbol{\iota}) \\ &\propto \prod_{n=1}^{N} \int_{\mathbb{R}^M} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D}_n) = f(\boldsymbol{d}_n)) d\boldsymbol{\xi}\end{aligned} \tag{4.153}$$

This ML equation is valid if conditional stochastic independence $\boldsymbol{Y} \perp f(\boldsymbol{D}) \,|\, \boldsymbol{\xi}$ holds true. This means that no DIF exists with respect to the function $f(\boldsymbol{D})$. This is no additional assumption, because it follows immediately from the general assumption given by Equation 2.60. Comparing Equation 4.153 with the MML equation of the B-MIRT model (see Equation 4.86) highlights the close relationship between the two models. There are two differences: (a) The item response propensities $P(D_i = y_i \,|\, \boldsymbol{\theta}; \boldsymbol{\phi})$ are not involved, and (b) the joint distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta})$ is replaced by the conditional distribution $g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D}_n) = f(\boldsymbol{d}_n))$. Hence, whereas $\boldsymbol{\theta}$ represent the information of $\boldsymbol{D}$ with respect to the

estimands $\boldsymbol{\iota}$ and $\boldsymbol{\xi}$ in the B-MIRT model, this information is replaced by $f(\boldsymbol{D})$ in the LRM model. Recall that if the missing data mechanism is nonignorable, then missingness is informative. It is essential that the information in $\boldsymbol{D}$ is sufficiently summarized in LRMs by finding the appropriate function $f(\boldsymbol{D})$. If such a function can be found, then the model of $\boldsymbol{D}$ can be left out and ML inference based on a conditional model of $\boldsymbol{Y}$ given $f(\boldsymbol{D})$ is sufficient. However, what is an appropriate function $f(\boldsymbol{D})$? This is easy to answer at the theoretical level; Using ML estimation procedures, the appropriateness of the function $f(\boldsymbol{D})$ is given if conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y_{mis}} \,|\, (\boldsymbol{Y_{obs}}, f(\boldsymbol{D}))$ holds true. In application, however, this is not testable. The best practice might be to find an appropriate model for $\boldsymbol{D}$ and to use summary measures containing the essential information, which most likely approximate the required conditional stochastic independence assumption. The use of sum scores S_D means that $\bar{D}$ or estimates $\hat{\boldsymbol{\theta}}$ are examples of such an approach.

In application of MML estimation in IRT models with latent regressions, distributional assumption needs to be made with respect to $g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D}))$. Typically, it is assumed that the M-dimensional latent residual $\boldsymbol{\zeta} = \zeta_1, \ldots, \zeta_M$ of the regression $E(\xi \,|\, f(\boldsymbol{D}))$ is multivariate normal with $\boldsymbol{\zeta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\zeta)$. The matrix $\boldsymbol{\Sigma}_\zeta$ is the variance-covariance matrix of the residual $\boldsymbol{\zeta}$. In the LRM, homogeneity of variance and covariances is assumed with respect to all dimension ζ_m, so that $Var(\zeta_m \,|\, f(\boldsymbol{D})) = Var(\zeta_m)$ and $Cov(\zeta_m, \zeta_{k \neq m} \,|\, f(\boldsymbol{D})) = Cov(\zeta_m, \zeta_{k \neq m})$.

Person parameter estimation in LRM for nonignorable missing data As in the case of the MIRT models, ML and WML person parameter estimates are not directly affected by the latent regression model. These estimates depend exclusively on the observed responses $\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}}$ and the item parameter estimates. Differences between the person parameter estimates between the model that ignores missing responses and the LRM follows from differences in item parameter estimates exclusively. This is not the case in Bayesian estimates such as the EAP and the MAP. Here, the information of background variables affects the individual posterior distribution of the $\boldsymbol{\xi}$ and the point estimates respectively. The EAP in the LRM for nonignorable missing data is given by

$$\hat{\xi}_{m;EAP} = \frac{\int_{\mathbb{R}} \xi_m \cdot \int_{\mathbb{R}^{m-1}} P(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D})) d\boldsymbol{\xi}}{\int_{\mathbb{R}^m} P(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D})) d\boldsymbol{\xi}}. \tag{4.154}$$

The prior is the conditional distribution $g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D}))$ instead of $g(\boldsymbol{\xi})$ in the simple model that ignores missing data. In general, if independent variables in a LRM are predictive of the latent ability then $g(\boldsymbol{\xi} \,|\, f(\boldsymbol{D})) \neq g(\boldsymbol{\xi})$. In this case the locations of the individual prior

distributions are the expected value $E(\boldsymbol{\xi} \mid f(\boldsymbol{D}))$ which can be different across test takers depending on the values $f(\boldsymbol{D}) = f(\boldsymbol{d})$. Hence, EAP estimates $\hat{\xi}_{m;EAP}$ shrink toward the expected values $E(\xi_m \mid f(\boldsymbol{D}))$ instead of $E(\xi_m)$. This should reduce the shrinkage effect of EAPs and increase the EAP-reliability. Comparing Equations 4.90 and 4.154, once more reveals the conceptual proximity of the B-MIRT model and the LRM. Both equations differ only in replacing $\boldsymbol{\theta}$ by $f(\boldsymbol{D})$. If the latent response propensity estimates are unbiased and sufficiently reliable the EAPs $\boldsymbol{\xi}_{\boldsymbol{EAP}}$ obtained from the LRM and the MIRT models for nonignorable missing data should be approximately equal.

Note that the estimate $\hat{\boldsymbol{\theta}}$ is also a function $f(\boldsymbol{D})$. This is counter-intuitive at first sight since any estimate is typically written as $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}$. In this case $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}$ is the measurement error. Similarly, the variable $\boldsymbol{D}$ of $\boldsymbol{\theta}$ can also be written as $\boldsymbol{D} = f(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_{\boldsymbol{D}}$. If the function $f()$ is correctly specified in real applications, all measurement error $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}$ in the estimate $\hat{\boldsymbol{\theta}}$ result from measurement error $\boldsymbol{\varepsilon}_{\boldsymbol{D}}$ in the response indicators. Hence $\hat{\boldsymbol{\theta}} = f(\boldsymbol{D})$. The estimates of the latent variable depend merely on the indicators in the measurement model, which in turn depend stochastically on the latent variable. From this point of view the general Equation of the LRM for missing responses (see Equation 4.140) holds also when the estimate $\hat{\boldsymbol{\theta}}$ is chosen as predictor in the LRM.

Application of the LRM to Data Example A The LRM for nonignorable missing data was also applied to Data Example A. Two LRMs were included with (a) the linear regression $E(\xi \mid S_D)$ and (b) the linear regression $E(\xi \mid \hat{\theta})$. For the latter, EAPs were obtained in a unidimensional measurement model of θ based on $\boldsymbol{D}$ alone. In a subsequent step, the item and person parameters of the measurement model of ξ based on $\boldsymbol{Y}_{\boldsymbol{obs}} = \boldsymbol{y}_{\boldsymbol{obs}}$ were estimated including the respective LRM. Altogether, four models were estimated using M*plus* 6 (Muthén & Muthén, 1998 - 2010) - two 1PL-LRMs and two 2PL-LRMs - since the 1PLM and the 2PLM were applied as measurement models of ξ. The bias of the estimates $\hat{\beta}_i$ and $\hat{\alpha}_i$ of the item difficulties and the item discriminations were analyzed. EAPs were chosen as person parameter estimates which were compared with the EAPs of the complete data model, the simple model that ignores missing data, and the B-MIRT Rasch model. The estimated standardized regression coefficients of the regression $E(\xi \mid S_D)$ were $\hat{b}_z = 0.706$ ($SE = 0.016$, $t = 44.503$, $p < 0.001$) in the 1PL-LRM, and were $\hat{b}_z = 0.706$ ($SE = 0.016$, $t = 44.307$, $p < 0.001$) in the 2PL-LRM. When the regression $E(\xi \mid \hat{\theta})$ were used in the LRM instead of $E(\xi \mid S_D)$,then the standardized regression coefficients were practically identical in both models (1PLM: $\hat{b}_z = 0.706$, $SE = 0.016$, $t = 44.781$, $p < 0.001$; 2PLM: $\hat{b}_z = 0.706$, $SE = 0.016$, $t = 44.523$, $p < 0.001$). The standardized

regression coefficient is equal to the correlation $Cor(\xi, \hat{\theta})$. Recall that Data Example A was simulated with a correlation $Cor(\xi, \theta) = 0.8$. The underestimation of the standardized regression coefficients reflects unreliability in both S_D and $\hat{\theta}$. The marginal reliability of the EAPs of $\hat{\theta}$ was $Rel(\hat{\theta}) = 0.871$. Accordingly, the attenuation corrected standardized regression coefficient is given by $0.706 \cdot \sqrt{0.871}^{-1} = 0.781$. This value is close to the true value $Cor(\xi, \theta) = 0.8$. Unfortunately, it is difficult to predict the effect of unreliability in $\hat{\theta}$ on bias reduction in item and person parameters, and a general answer cannot be given here. The effect was only studied empirically using Data Example A.

The mean bias of the estimated item difficulties $\hat{\beta}_i$ was 0.057. This is not significantly different from zero ($t = 1.931, df = 29, p = 0.063$). Furthermore, the regression coefficient of $E(\hat{\beta}|\beta)$ is not significantly different from one ($Slope = 1.011$, $t = 0.472$,$SE = 0.023$, $p = 0.637$), implying that the bias is independent of the true item difficulties. Recall, in the unidimensional model ignoring missing data the bias of $\hat{\beta}_i$ was uncorrelated with the estimand[23]. The MSE of the estimates $\hat{\beta}_i$ was 0.016. This is exactly the same value as found in the MIRT models applied to Data Example A (see Section 4.5.3.2). Figure 4.28 shows the estimated item difficulties from the both one-parameter LRMs including either $E(\xi|S_D)$ or $E(\xi|\hat{\theta})$ respectively, compared to the true values β_i. Apparently, the estimates of both models are practically identical. Furthermore, in Figure 4.29 (left graph), the equality between the estimated item difficulties of the LRM and the B-MIRT model is shown. In the right graph of Figure 4.29, it can be seen that the increased underestimation of item difficulties if missing data are ignored was corrected using the LRM. The 2PL-LRM was also applied to study the estimation of the discrimination parameters in the LRM. The simulation study reported in Section 3 revealed that, on average, the estimation of α_i is not systematically biased. So, the focus here is on the comparison between the estimates $\hat{\alpha}_i$ of the different models applied to the data. The estimates were, on average, unbiased. The mean of the estimated discrimination parameters was $\bar{\hat{\alpha}} = 1.017$ in the LRM using $E(\xi|S_D)$ and $\bar{\hat{\alpha}} = 1.014$ with the regression $E(\xi|\hat{\theta})$. This is not significantly different from one in both cases (2PL-LRM with $E(\xi|S_D)$: $Bias = 0.017$, $t = 0.786$, $df = 29$, $p = 0.438$; 2PL-LRM with $E(\xi|\hat{\theta})$: $Bias = 0.014$, $t = 0.648$, $df = 29$, $p = 0.522$). Figure 4.30 shows that the estimates $\hat{\alpha}_i$ of both LRMs are very close, which is also reflected by the similar mean squared errors of $MSE = 0.015$ with $E(\xi|S_D)$ and $MSE = 0.014$ with $E(\xi|\hat{\theta})$. This is close to the values of the MIRT models applied to Data Example A. Accordingly, Figure 4.31 illustrates that the estimated item

[23] Recall that in the simple model that ignores missing data the bias was dependent on the true item difficulties (see Section 3.2.2).

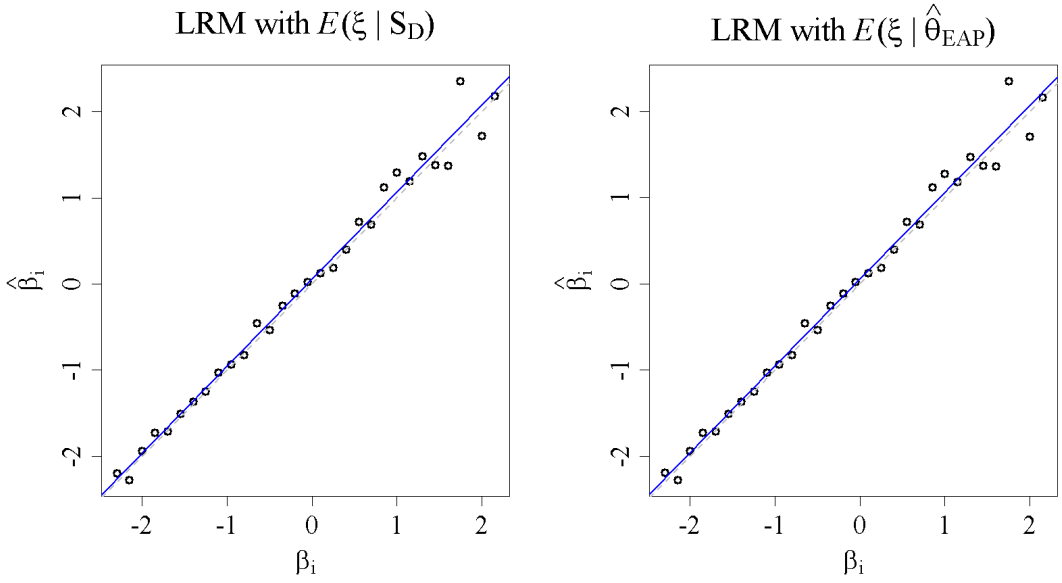

Figure 4.28: Estimated item difficulties in the 1PLM including the latent regression model with $E(\xi \mid S_D)$ (left) and $E(\xi \mid \hat{\theta})$ (right). The grey dotted lines indicate bisectric lines. The blue lines are regression lines.

discriminations of the LRM with $E(\xi \mid S_D)$ and the 2PL-BMIRT model differ only negligibly. Finally, the EAP estimates from the different models were compared [24]. Figure 4.32 compares the EAPs of different models including the two 1PL-LRMs with either $E(\xi \mid S_D)$ or $E(\xi \mid \hat{\theta})$. Not only are the EAPs of the two LRMs almost equal, but the correlation with the EAPs obtained using the B-MIRT Rasch model was very close to one as well. It can be seen that the bias toward the mean, especially in the lower range of ξ, was considerably reduced in both the 1PL-LRMs and the 1PL-B-MIRT model. An identical pattern was found in for the EAPs of the 2PL-LRM and the 2PL-BMIRT model. Therefore, a detailed presentation of the results was renounced.

Model equivalence In Section ,three criteria were introduced to judge equivalence with respect to IRT models for (non)ignorable missing data. These are (a) equivalence in the construction in the latent variable $\boldsymbol{\xi}$, (b) equivalence in bias reduction of item and person parameter estimates, and (c) same model fit. If MIRT models for nonigorable missing data and LRMs including $f(\boldsymbol{D})$ are compared with respect to these criteria, then the two

[24] As in the case of MIRT models, the WML and ML estimates are hardly affected by the LRM and have been left out here.

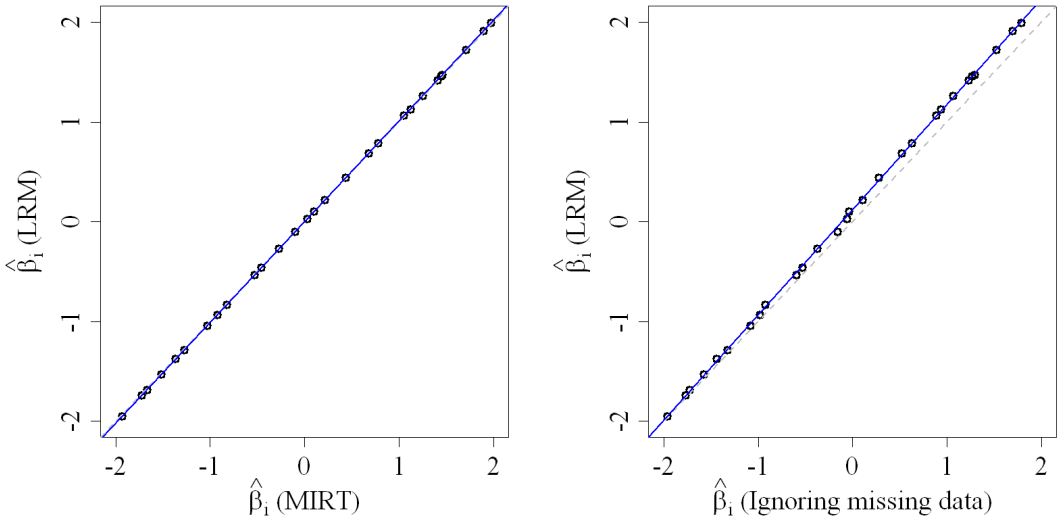

Figure 4.29: Comparison of item difficulty estimates obtained by the 1PL-LRM, with the regression $E(\xi \mid S_D)$, with the BMIRT Rasch model (left), and the unidimensional IRT model ignoring missing data (right). The grey dotted lines represent the bisectric. The blue lines are regression lines.

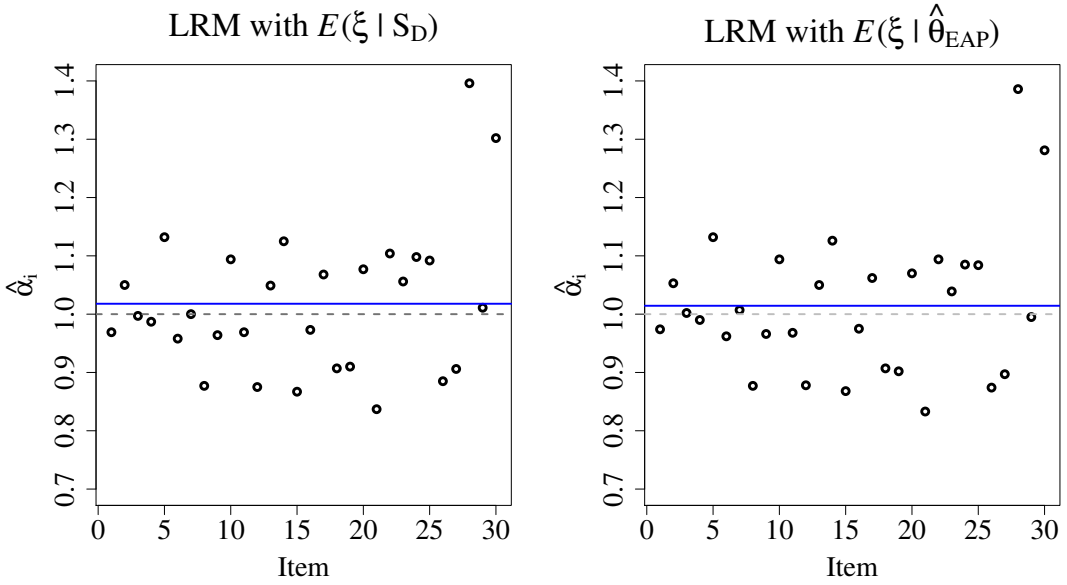

Figure 4.30: Estimated item discriminations in the 2PLM including the latent regression model with $E(\xi \mid S_D)$ (left) and $E(\xi \mid \hat{\theta})$ (right). The grey dotted line indicates the true value $\alpha_i = 1$ and the blue line indicates the mean $\bar{\hat{\alpha}}_i$

approaches turned out to be equivalent with respect to (a). The target model, here the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ is equally preserved in both - MIRT models and LRMs. Furthermore, the two approaches are also equivalent with respect to the bias reduction parameter estimates if three conditions are met. First, the assumptions of the MIRT model must hold true, especially the conditional stochastic independence assumptions (see Equations 4.5.4 and 4.74). Second, an appropriate function $f(\boldsymbol{D})$ must be found so that conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y_{mis}} \mid (\boldsymbol{Y_{obs}}, f(\boldsymbol{D}))$ is met. Third, the regression $E[\boldsymbol{\xi} \mid f(\boldsymbol{D})]$ must be correctly specified. The stronger the violation of the conditional stochastic independence assumption , and the more the latent regression is misspecified, the stronger the lack of equivalence in the bias reduction. In turn, the LRM potentially outperform MIRT models in the bias reduction of parameter estimates if certain assumptions of the MIRT models are not met. For example, if the latent ability dimensions ξ_m and the latent response propensities θ_l are non-linearly related, then the MIRT model can fail to adjust the bias. The LRM, however, allows for multiple polynomial regressions based on the estimates $\hat{\theta}_l$. The question regarding which approach - MIRT model or LRM - should be preferred needs to be answered in accordance to the particular application.

If $f(\boldsymbol{D}) \neq \boldsymbol{D}$, then MIRT models and LRMs are neither nested nor do they include

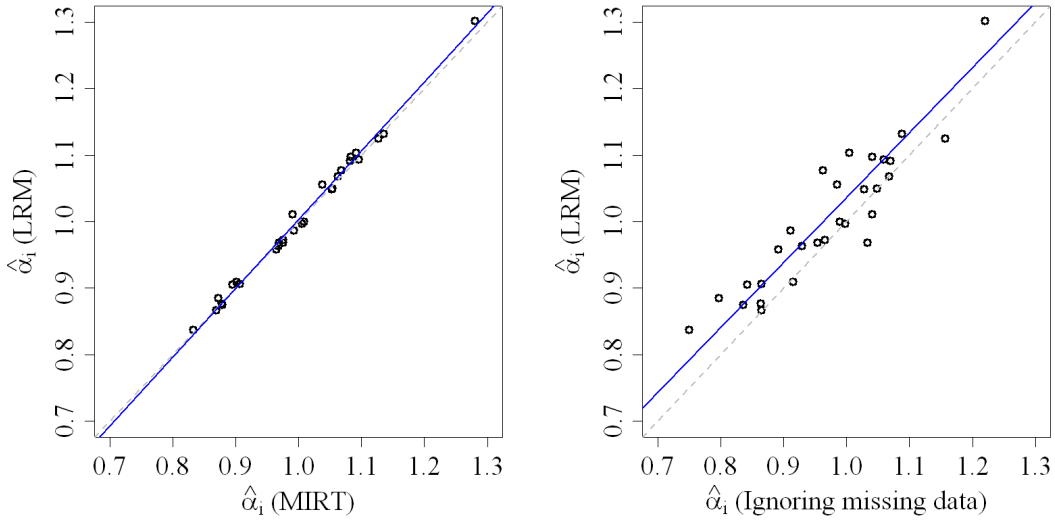

Figure 4.31: Comparison of the estimated item discriminations of the 2PL-LRM with $E(\xi \mid S_D)$ and the 2PL-BMIRT model (left), and the unidimensional IRT model ignoring missing data (right). The grey dotted lines denote the bisectric. The blue lines are regression lines.

the same variables. Therefore, it is difficult to judge equivalence in terms of model fit. Information criteria are the only measures to compare MIRT models and LRMs. However, depending on the chosen function $f(\boldsymbol{D})$ and the specification of the latent regression, LRMs can be much more parsimonious than MIRT models. Instead of the parameters of the measurement model of $\boldsymbol{\theta}$, only the parameters of the latent regression need to be estimated. For that reason, information criteria might tend to favour LRMs. Of course, this does not imply that LRMs are the better choice to adjust for nonignorable missing data. Model fit criteria are not sensitive to the bias correction. Differences highlight only that MIRT models and LRMs are not equivalent in terms of model fit.

Extentions of the LRM for nonignorable missing data Further extensions of the LRM are possible and in some applications even required. For example, let there be different test booklets as in many large scale assessments. In PISA or NEAP, a balanced incomplete block design was chosen so that each student only answered a small portion of the complete item pool (e. g. von Davier et al., 2006; D. Li et al., 2009). It is nearly impossible to create equivalent test booklets: Differences in the stimulus material such as different text length might occur. Or it might happen that the booklets vary with respect to the average

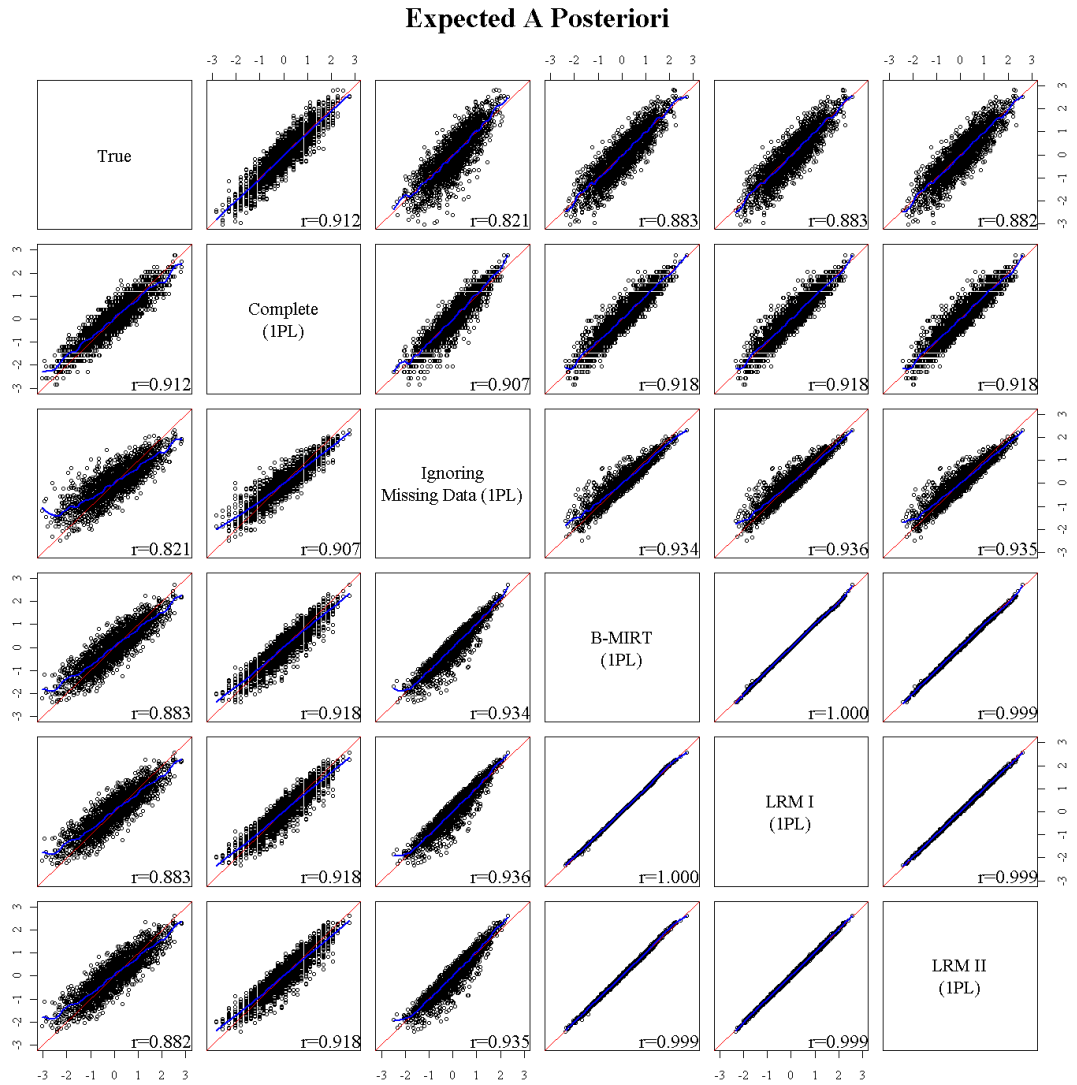

Figure 4.32: Comparison of the true values of ξ underlying Data Example A with the respective EAP person parameter estimates obtained from different models including the LRM I with $E(\xi \mid S_D)$ and LRM II with $E(\xi \mid \hat{\theta})$. The red lines represent the bisectric. The blue lines are smoothing spline regressions.

of the items difficulties. This can result in different distributions of $f(\boldsymbol{D})$ and can lead to interaction effects between the test booklet and $\boldsymbol{D}$ with respect to $\boldsymbol{Y}$ and $\boldsymbol{\xi}$, respectively. More formally, let there be k booklets. A vector $\boldsymbol{I_B} = I_{B=1}, \ldots, I_{B=k}$ of indicator variables of the single booklets B can be created. $\boldsymbol{I_B}$ can moderate the regressive dependency between $\boldsymbol{\xi}$ and $f(\boldsymbol{D})$. This can easily be taken into account using conditional regressions $E(\boldsymbol{\xi} \mid f(\boldsymbol{D}), \boldsymbol{I_B}) = f_0(\boldsymbol{I_B}) + f_1(\boldsymbol{I_B}) \cdot f(\boldsymbol{D})$ that allow for interaction effects. Alternatively, a

multiple group IRT-LRM can be used, with the booklet as the grouping variable. The parameters of the conditional regression of $\boldsymbol{\xi}$ on $f(\boldsymbol{D})$ are allowed to vary across the groups (booklets). Of course, this is only one specific example underlining the importance as well as the flexibility of the correct inclusion of $f(\boldsymbol{D})$ in the latent regression. It is important to note that the correct specification of the latent regression with $f(\boldsymbol{D})$ is inevitable to account for missing responses properly. Other extensions might also be plausible or required in order to account for missing data depending on the study design and other factors that need to be considered in a particular application. The major advantage of LRMs is their flexibility that allows to include additional variables and interaction terms that reflect the complexity of the study and the design. Unfortunately, many commonly used IRT software packages such as BILOG-MG (Zimowski et al., 1996), PARSCALE (Muraki & Bock, 2002), or MULTILOG (du Toit, 2003) do not allow for the inclusion of a LRM. Multiple Group IRT (MG-IRT) models for nonignorable missing data might be a solution, if a discrete function $f(\boldsymbol{D})$ can be found. This approach is discussed in the next section.

4.5.5 Multiple Group IRT Models for Nonignorable Missing Data

The multiple group IRT (MG-IRT) models for nonignorable missing data are discussed here as a special case of latent regression models for nonignorable missing data introduced in the previous section. Rose, von Davier, and Xu (2010) came up with the idea to account for nonignorable missing data by stratification of $\bar{D}$. They reanalyzed the PISA 2006 data and used a multiple group model including three strata of $\bar{D}$ which referred to test takers with low, medium, and high proportions of missing responses. In the previously introduced terminology, the stratified variable $\bar{D}$ is also a function $f(\boldsymbol{D})$ that can be used either as a predictor in a LRM or, alternatively, as a grouping variable in a MG-IRT model. Let $X = f(\boldsymbol{D})$ be a categorical variable that serves as a grouping variable in the MG-IRT model. Considering ML-estimation in LRM for nonignorable missing responses, the conditional stochastic independence assumption $\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, (\boldsymbol{Y}_{obs}, f(\boldsymbol{D}))$ was found to be sufficient to account for missing data that are NMAR. Since the MG-IRT model is conceptually equivalent to the LRM in cases of discrete functions $X = f(\boldsymbol{D})$, conditional stochastic independence

$$\boldsymbol{D} \perp \boldsymbol{Y}_{mis} \,|\, (\boldsymbol{Y}_{obs}, X) \tag{4.155}$$

is assumed respectively.

ML estimation in MG-IRT models for nonignorable missing data A detailed derivation of the ML estimator of the MG-IRT model is renounced here, due to theoretical equivalence of LRM and MG-IRT models. Since X is a discrete function of the response indicator vector $\boldsymbol{D}$, the term $f(\boldsymbol{D})$ in Equation 4.153 needs to be replaced by X to yield the MML estimation function of the MG-IRT model for non-ignorable missing data given by

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{x}; \boldsymbol{\iota}) \propto \prod_{x=1}^{G} \prod_{n_x=1}^{N_x} \int_{\mathbb{R}^M} g(\boldsymbol{Y_{n;obs}} = \boldsymbol{y_{n;obs}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \,|\, X_n = x_n; \boldsymbol{\iota}) d\boldsymbol{\xi}, \tag{4.156}$$

with G as the number of groups and N_x as the sample size in group $X = x$. Hence, $N = \sum_{x=1}^{G} N_x$. If local stochastic independence $Y_i \perp Y_j \,|\, \boldsymbol{\xi}$ (for all $i \neq j$) holds true, then Equation 4.156 can be written as

$$\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{x}; \boldsymbol{\iota}) \propto \prod_{x=1}^{G} \prod_{n_x=1}^{N_x} \int_{\mathbb{R}^M} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} g(\boldsymbol{\xi} \,|\, X_n = x_n; \boldsymbol{\iota}) d\boldsymbol{\xi}. \tag{4.157}$$

The conditional distributions $g(\boldsymbol{\xi} \,|\, X = x; \boldsymbol{\iota})$ are typically assumed to be multivariate normal with

$$\boldsymbol{\xi} \,|\, X = x \sim N[E(\boldsymbol{\xi} \,|\, X = x), \Sigma_{\boldsymbol{\xi} \,|\, X=x}]. \tag{4.158}$$

Hence, the variance-covariance matrices can vary across groups $X = x$.

Comparison between the LRM with $E(\boldsymbol{\xi} \,|\, X)$ and the MG-IRT model If X has H values, then a single group IRT model with a LRM using $H - 1$ indicator variables $I_{X=x}$ is conceptually equivalent to a MG-IRT model with H groups. However, in typical implementations of latent regression models in IRT software, such as M*plus* (Muthén & Muthén, 1998 - 2010) or ConQuest (Wu et al., 1998), variances and variances are assumed to be equal. That is, only *one* variance-covariance matrix Σ_{ζ} of the residual $\boldsymbol{\zeta} = \zeta_1, \ldots, \zeta_M$ is estimated across the groups in the LRM. If the variance-covariance structure is identical in all groups x of X, then $\Sigma_{\zeta} = \Sigma_{\xi \,|\, X=x}$, for all $x = 1, \ldots, H$. Therefore, the MG-IRT model is less restrictive and might be preferred provided that an appropriate discrete variable $X = f(\boldsymbol{D})$ can be found. Furthermore, in the LRM it is implicitly assumed that no DIF exists with respect to $f(\boldsymbol{D})$. In MG-IRT models the item parameters are explicitly constraint to be equal across the groups x of X to establish a common metric in all groups.

It should be noted that each missing pattern $\boldsymbol{D} = \boldsymbol{d}$ could be considered a group in

an MG-IRT model. This model is equivalent to a pattern mixture model with certain assumptions such as measurement invariance with respect to $\boldsymbol{D}$. Unfortunately, there are different problems with this approach. In all groups, the single items Y_i are either completely observed or completely missing. Furthermore, there are theoretically 2^I missing patterns. Hence, in cases with a realistic number of items, the sample size needs to be large in order to have sufficient numbers of cases for each observed missing pattern.

Person parameter estimation in the MG-IRT models for nonignorable missing data ML and WML person parameter estimation depends exclusively on the observed responses $\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}}$ and the item parameter estimates. Therefore, the bias reduction in person parameter estimates rests upon bias reduction in item parameter estimates. In contrast, EAP person parameter estimation allows one to take additional information into account. For example, informative background variables can be included in an LRM. The term informative variables refers to variables that are stochastically dependent on the estimand $\boldsymbol{\xi}$. Recall that $\boldsymbol{D}$ is informative regarding $\boldsymbol{\xi}$ and the item parameters in the case of nonignorable missing data. For that reason functions $f(\boldsymbol{D})$ are used in LRMs for nonignorable missing responses. In the MG-IRT model for item nonresponses the grouping variable X is a discrete function $f(\boldsymbol{D})$. The group membership expressed by the values x of X is informative with respect to item end person parameters and is, therefore, taken into account in EAP estimation. In technical terms, that means that each group $X = x$, has their own prior distribution $g(\boldsymbol{\xi} \mid X = x)$ of the latent variable. Generally, the EAP in the MG-IRT model is defined as

$$\hat{\xi}_{m;EAP} = \frac{\int_{\mathbb{R}} \xi_m \cdot \int_{\mathbb{R}^{m-1}} P(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \mid \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \mid X = x) d\boldsymbol{\xi}}{\int_{\mathbb{R}^m} P(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}} \mid \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi} \mid X = x) d\boldsymbol{\xi}}. \tag{4.159}$$

Recall that in the simple unidimensional model that ignores missing data, the unconditional distribution $g(\boldsymbol{\xi})$ is taken as the prior distribution. If $X \not\perp \boldsymbol{Y}$, which is implied by $X \not\perp \boldsymbol{\xi}$ under local stochastic independence, then the differences in the latent proficiency levels between persons with different missing patterns expressed by group membership x of X are taken into account by the priors $g(\boldsymbol{\xi} \mid X = x)$. As a consequence, the shrinkage is reduced since the EAPs shrink toward the expected values $E(\xi_m \mid X = x)$ instead of the unconditional means $E(\xi_m)$. The stronger the stochastic relation between X and $\boldsymbol{\xi}$, the more informative the missingness with respect to the estimand $\boldsymbol{\xi}$ is, and the more the shrinkage effect is reduced in the MG-IRT model compared to ignoring missing responses.

Model equivalence The MG-IRT model is a special case of the LRM for nonignorable missing data. Accordingly, the issue of model equivalence is analogous to the LRM as well (see page 232). In summary, the MG-IRT model is equivalent with respect to the construction of the latent variable $\boldsymbol{\xi}$, implying that the item parameters are also equivalent to the target model. The MG-IRT model is not expected to be equivalent to B-MIRT and LRM models with continuous functions $f(\boldsymbol{D})$. However, if an appropriate function $X = f(\boldsymbol{D})$ can be found that preserves the essential information of $\boldsymbol{D}$ with respect to the estimands in the target model, then the bias reduction will be close to that of B-MIRT models and LRMs for missing responses that are NMAR. Looking in more detail at the the sufficient condition $\boldsymbol{D} \perp \boldsymbol{Y_{mis}} \,|\, (\boldsymbol{Y_{obs}}, X)$ underlying MG-IRT models reveals that this assumption will most hold if the number of groups is large or the latent response propensity is discrete. Indeed, at least theoretically, $\boldsymbol{D}$ can result from latent classes that refer to typical missing patterns. To use latent class analysis to model $\boldsymbol{D}$ is not considered here. However, this approach is theoretically close to pattern mixture models and potentially worthwhile to pursue. If a continuous latent response propensity $\boldsymbol{\theta}$ exists, it might be difficult to find an appropriate discrete function $f(\boldsymbol{D})$ that can serve as a grouping variable in an MG-IRT model. In such cases, the MG-IRT model is likely to reduce the bias less compared to MIRT models or LRMs with continuous functions. MG-IRT models include different variables than MIRT models. As LRMs, multiple group models are difficult to compare with MIRT models in term of model fit. Since the measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D}$ is not included in the MG-IRT model, the latter is typically much more parsimonious unless the number of groups is extremely high. If information criteria are used to compare MIRT and MG-IRT models, then more parsimonious models are typically preferred. Recall that this does not mean that the more parsimonious model accounts better for missingness.

MG-IRT models as an alternative to high-dimensional MIRT models for nonignorable missing data For the reanalysis of the data of PISA 2006, Rose, von Davier, and Xu (2010) simply created three strata based on $\bar{D}$. This approach might be justifiable if a unidimensional latent variable θ can be constructed based on $\boldsymbol{D}$. As outlined in the previous section, $\bar{D}$ or S_D can be seen as manifest test scores that are increasingly correlated with θ when the number of items increases. Hence, the grouping variable for the MG-IRT model that is generated using $\bar{D}$ or S_D is constructed by fallible measures of θ. The situation becomes difficult if $\boldsymbol{\theta}$ underlying $\boldsymbol{D}$ is multidimensional, especially with low correlations $Cor(\theta_l, \theta_k)$. If between-item multidimensional holds for the measurement model

of $\boldsymbol{\theta}$, then groups can be formed as combinations of all stratified variables S_{Dl}, where S_{Dl} is the sum of only those response indicators D_i that constitute the measurement model of θ_l. Additionally, if within-item multidimensionality exists in the measurement model of $\boldsymbol{\theta}$,then the use of S_{Dl} is critical. Alternatively, the estimates $\hat{\theta}_l$ can be estimated in a first step fitting an MIRT model to $f(\boldsymbol{D})$. In a second step, the combinations of all stratified estimates $\hat{\theta}_l$ can be used as a grouping variable in the MG-IRT model. This approach is recommended if LRMs are not available. This approach avoids the use of high dimensional MIRT models and can also reduce the missing-related bias substantially. The determination of the number of groups might depend on several factors such as the sample size, the number of dimensions θ_l, and the desired accuracy. The more fine-grained the stratification, the more precise is the adjustment of the bias due to missing data. Fortunately, the empirical results of Rose et al. (2010) suggest that stratification can be pretty rough. They used only three strata and yielded nearly identical results compared to the between-item multidimensional IRT model. This will be demonstrated next, applying the MG-IRT model to Data Example A.

Application of the MG-IRT model to Data Example A In Data Example A, the latent response propensity is known to be unidimensional. The sum score S_D was used to form groups. Three strata were determined in such a way that the resulting groups are similar with respect to the group sizes. Group 1 consisted of $n_1 = 676(33.8\%)$ with 13 or less answered items. Test takers with 14 - 17 completed items were in group 2 with $n_2 = 722(36.1\%)$. Group 3 consisted of cases with more than 18 item responses ($n_3 = 602(30.1\%)$). Two MG-IRT models were applied: The MG-IRT Rasch model (1PL-MG-IRT model) and the MG-IRT Birnbaum (2PL-MG-IRT model). The item and person parameter estimates were compared with the true values underlying Data Example A and with the respective estimates of the MIRT models and LRMs for nonignorable missing data. M*plus* was used for parameter estimation. The input file is given in Listing A.8 in Appendix 5.3. In order to obtain comparability of the estimates from the different models, the expected value $E(\xi)$ over the groups was fixed to one using nonlinear constraints. Hence, the weighted sum of the three group means $E(\xi \mid X = x)$ was set to zero. [25]. The distribution of the ξ differs considerably. The estimated means were $\bar{\xi}_1 = -0.735$ ($s_1^2(\xi) = 0.579$) in group one, $\bar{\xi}_2 = 0.007$ ($s_2^2(\xi) = 0.507$) in group two, and $\bar{\xi}_3 = 0.816$ ($s_3^2(\xi) = 0.599$) in the third group. Due to Cohen's d, the effect sizes of the pairwise mean

[25] $E(\xi) = E[E(\xi \mid X)] = \sum_{x=1}^{3} P(X = x)E(\xi \mid X = x)$. The probabilities $P(X = x)$ were replaced by the relative frequencies of the three groups.

differences were large. Using the pooled standard deviation ($s_{pool}(\xi) = 0.748$) to determine Cohen's $d_{xx'}$ between group $X = x$ and $X = x'$, the effect sizes were $d_{21} = 0.991$, $d_{32} = 1.081$, and even $d_{31} = 2.074$. This reflects the strong dependency between the proportion of missing data and the underlying variable ξ. Large effect sizes were also found in real data analyses. For example, Rose et al. (2010) applied the MG-IRT model to the PISA 2006 data with the stratified response rate as grouping variable. They reported effect sizes of $d_{xx'} \approx 1$ in the mean differences of the latent variables. These differences in the latent ability distribution between groups of different proportions of missing data are taken into account in the parameter estimation in MG-IRT models. This corrects for nonignorable missing responses. In turn, if the strata do not vary with respect to the distribution of ξ or $\boldsymbol{\xi}$, in the multidimensional case, and the assumption of conditional stochastic independence given by Equation 4.155 hold then the missing data mechanism is ignorable.

The estimated item difficulties obtained by the 1PL-MG-IRT model were compared with the true item difficulties. Figure 4.33 reveals that $\hat{\beta}_i$ from the 1PL-MG-IRT model and the B-MIRT model are nearly identical. A comparison with the true values β_i shows that the systematic bias found in unidimensional model of ξ that ignores missing data has vanished. Accordingly, the slope of the regression of the estimates $\hat{\beta}$ on the true item diffulties was not significantly different from one ($slope = 0.970$, $SE = 0.017$, $t = -1.742$,$p = 0.174$), and the intercept was very close to zero ($intercept < 0.001$, $SE = 0.022$, $t = 0.027$,$p > 0.999$). Accordingly, the mean bias of the estimates $\hat{\beta}_i$ in the MG-IRT model is 0.004. This is also not significantly different from zero ($t = 0.179$, $df = 29$, $p = 0.859$). The mean squared error was $MSE = 0.016$ and, therfore, exactly the same as in the B-MIRT Rasch model. In the lower two graphs of Figure 4.33 the estimates $\hat{\alpha}_i$ of the item discriminations are shown. As expected, the estimates of the 2PL MG-IRT model and the 2PL B-MIRT model are very similar. The mean item discrimination was $\bar{\hat{\alpha}} = 1.014$. This is not significantly different from one ($t = 0.633$, $df = 29$, $p = 0.532$). The mean squared error was the same as in the 2PL B-MIRT model ($MSE = 0.014$).

Finally, the EAP estimates have been compared with the true values of ξ and EAPs obtained with other IRT models applied to Data Example A. Figure 4.34 summarizes the results. The colors black, red, and blue in Figure 4.34 mark the three strata of S_D which served as grouping variables in the MG-IRT model. The ellipsoids in the upper left graph are drawn so that all cases pertaining to the respective group are inside. The correlation between ξ and the EAPs from the MG-IRT model was $r(\xi, \hat{\xi}_{EAP}) = 0.867$. This is slightly lower than in the 1PL-B-MIRT model ($r(\xi, \hat{\xi}) = 0.883$) and the 1PL-LRM ($r(\xi, \hat{\xi}) =$

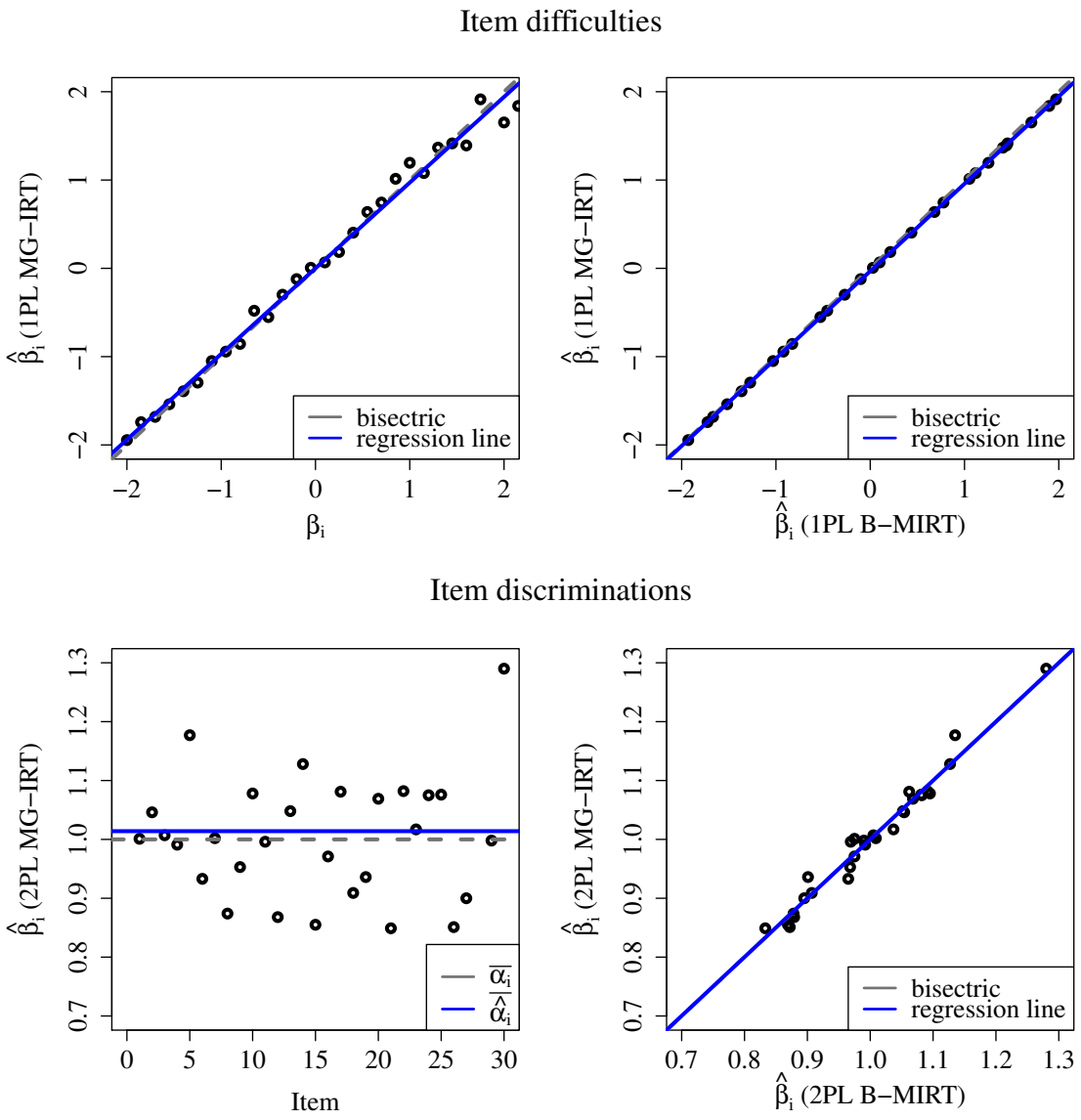

Figure 4.33: True and estimated item difficulties of the MG-IRT and the 1PL-BMIRT model (upper row), and true and estimated item discriminations of the 2PL-MG-IRT and the 2PL-BMIRT model (lower row) using Data Example A.

0.882). This illustrates the distributional differences of ξ across the strata for both the true values of ξ and the EAP estimates. The bias reduction of the EAP estimates becomes

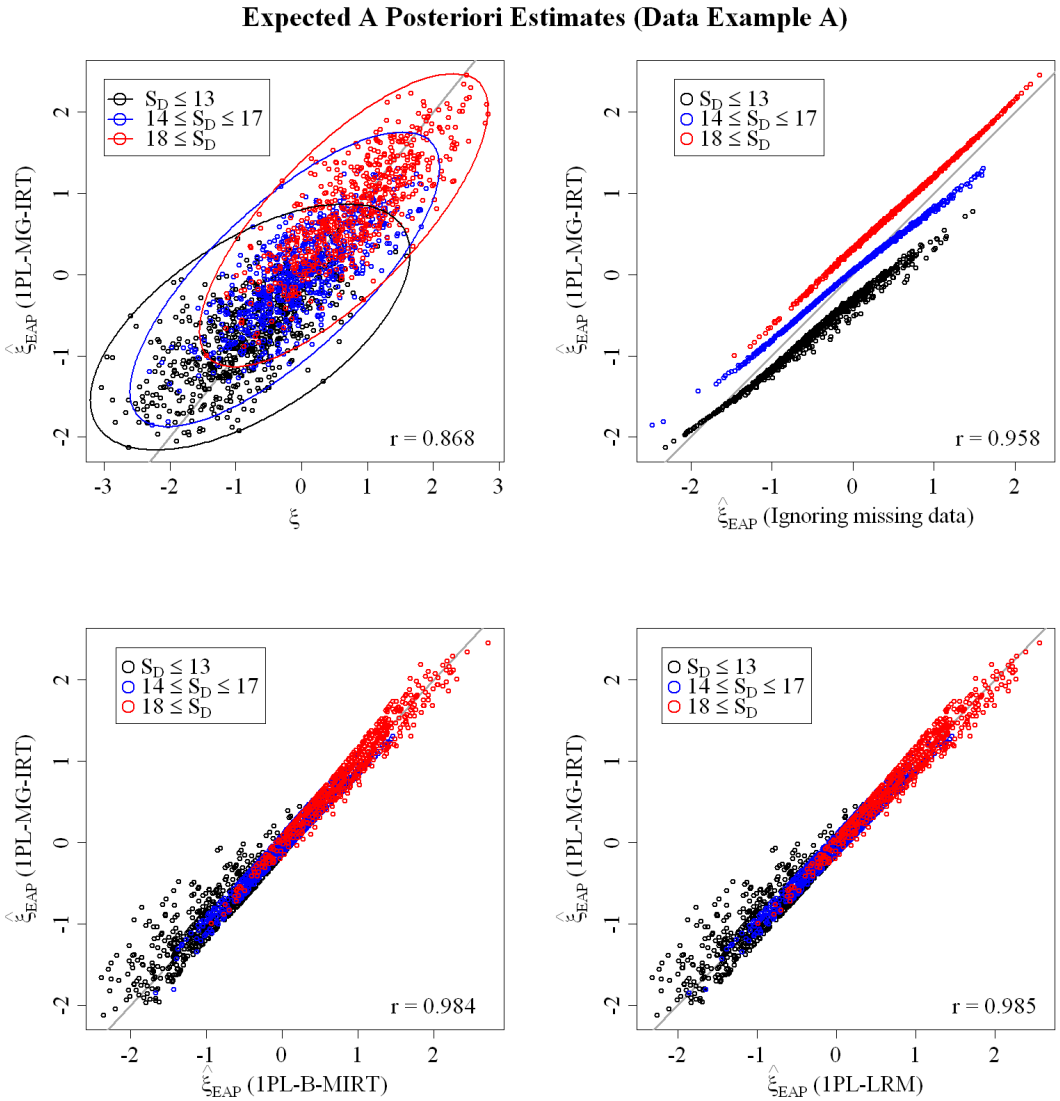

Figure 4.34: EAP estimates from the 1PL-MG-IRT model compared with the true values of ξ (upper left), and the EAP estimates from alternative models applied to Data Example A. The grey lines represent the bisectric.

obvious in the upper right graph of Figure 4.34. Here, the estimates of two models, the unidimensional 1PLM that ignores missing data and the 1PL-MG-IRT model, are plotted. The variance $s^2(\hat{\xi}_{EAP}) = 0.632$ in the model that ignores missing data is considerably lower than $s^2(\hat{\xi}_{EAP}) = 0.706$ in the MG-IRT model. As explained previously, this is due

to including the conditional distribution $g(\xi \mid X = x)$ in the EAP estimation. Since $X = f(\boldsymbol{D})$, the distributional differences of ξ given $\boldsymbol{D}$ are taken into account. The additional information of $\boldsymbol{D}$ with respect to ξ is reflected by the reduced shrinkage effect. The EAPs tend toward the respective exptected value $E(\xi \mid X = x)$ instead of the unconditional expected value $E(\xi)$.

In the introduction of this section it was argued that the LRM and the MG-IRT model are conceptually equivalent. Both rest upon the inclusion of functions $f(\boldsymbol{D})$. In the MG-IRT model these functions have to be discrete, whereas (quasi-)continuous[26] functions can be used in the LRM. Therefore, the correlation between the EAPs of the 1PL-LRM model and the 1PL-MG-MIRT model is $r = 0.985$. However, the impact of the categorization of S_D can be seen graphically in the lower two graphs of Figure 4.34. Recall that the correlation between the EAPs of the B-MIRT Rasch model and the 1PL-LRM was $r > 0.999$. Obviously, the use of a roughly categorized function of S_D lowers the correlation with the true latent variable ξ as well as with the EAP estimates from the MIRT models and the LRM. On average, the effect seems negligible, but at the individual level the differences may be substantial for some cases. The largest difference between the EAP estimates of the 1PL-LRM and the 1PL-MG-IRT model in Data Example A was 0.376. Considering that the standard deviation of ξ within the strata is on average $s_{pool}(\xi) = 0.747$, this difference corresponds to half a standard deviation. Especially in cases of the strata with a maximum of 13 answered items, non-negligible differences between the EAP estimates occurred. Insofar, the LRM and the MIRT models seem to be superior to the MG-IRT model with respect to Bayesian person parameter estimates at the individual level.

4.5.6 Joint Modelling of Omitted and Not-reached Items

So far, in this work differences in item nonresponses resulting from not-administered items, omissions, and not-reached items at the end of the test have not been addressed in detail. However, these differences have implications regarding the suitability of the different model-based approaches that were examined in the previous sections. Planned missing data result from not-administered items due to the item design, such as balanced incomplete block design or multi-matrix sampling (Frey et al., 2009; Van der Linden, Veldkamp, & Carlson, 2004). Since planned missing data are typically MCAR, they are

[26] $\boldsymbol{D}$ is a discrete variable with 2^I values - the missing patterns. Hence, strictly speaking, the functions $\boldsymbol{D}$ are always discrete. However, if the function $f(\boldsymbol{D})$ has a large number of possible values, then it can be treated as continuous variable in a LRM.

not further considered here. However, they need to be distinguished from omitted and not-reached items. Note that if the booklets are randomly assigned to test takers in a multi-matrix sampling design, then planned missing data due to not-administered items are stochastically independent of the person variable U and, therefore, of any function $f(U)$ such as the latent ability $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. However, missingness due to omitted or not-reached items are potentially related to the U and $(\boldsymbol{\xi}, \boldsymbol{\theta})$ respectively. If $\boldsymbol{D}$ is used as an indicator of a latent response propensity in IRT models for missing responses, then the indicators D_i should only indicate the responses or nonresponses of the items actually administered to the respective test taker. Otherwise, D_i should be regarded as missing as well. In this case, it is ensured that $\boldsymbol{D}$ is an indicator of a person?s tendency to respond to test items not confounded by information of test design independent of the test takers. The remaining question is whether missing responses due to omitted or not reached items can be treated equally or not. This question will be answered in the remainder of this section.

4.5.6.1 Differences Between Omitted and Not-reached Items

In both cases - omitted and not-reached items - the resulting missing responses w.r.t. Y_i can be MCAR, MAR, or NMAR. However, there is some empirical evidence that the probability of omissions and the probability not to reach the end of the test are related. Culbertson (2011, April) found that the tendency to omit items increases with lower proficiency levels, whereas the probability of not reaching the end of the test decreases with lower ability levels. Possibly, test takers with high omission rates reach the end of the test faster. Hence, the more omitted responses, the less not-reached items. Especially in timed tests, such relations can be expected. In such cases, it seems inappropriate to handle omitted and not-reached items equally. For example, it seems suitable to assume a single latent response propensity in a B-MIRT model for nonignorable missing responses is inconsistent with a negative correlation between the probability of omissions and the probability not to reach the end of the test. Apart from empirical evidence suggesting different treatments of omitted items and not-reached items, there are important formal differences.

To illustrate the difference between missing data due to omissions of items and missing data due to not-reached items, a small data example $\boldsymbol{D} = \boldsymbol{d}$ with $N = 40$ test takers and $I = 10$ items was simulated. Three conditions were considered: (a) missing responses resulting from not reached items, (b) missing responses due to omissions, and (c) item nonresponses due to omissions of items *and* failing to reach the end of the test. The resulting

indicator matrices $\boldsymbol{d}$ with the missing data patterns are presented graphically in Figure 4.35. The persons are ordered according to their number of reached items. The items are ordered with respect to their position in the test. If missing responses occur solely due to not-reached items, then the response indicator matrix shows a perfect Guttman pattern (Andrich, 1985; Guttman, 1950). In terms of missing data theory, this is a monotone missing pattern (Little & Rubin, 2002; McKnight et al., 2007) that is often found in longitudinal studies due to attrition over time. In contrast, the second graph in Figure 4.35 gives the missing data pattern when the time to complete the test was unlimited. Hence, all test takers completed the test and missing responses resulted only from omissions. In this case, the pattern of the indicator matrix is non-monotone. Interestingly, the different

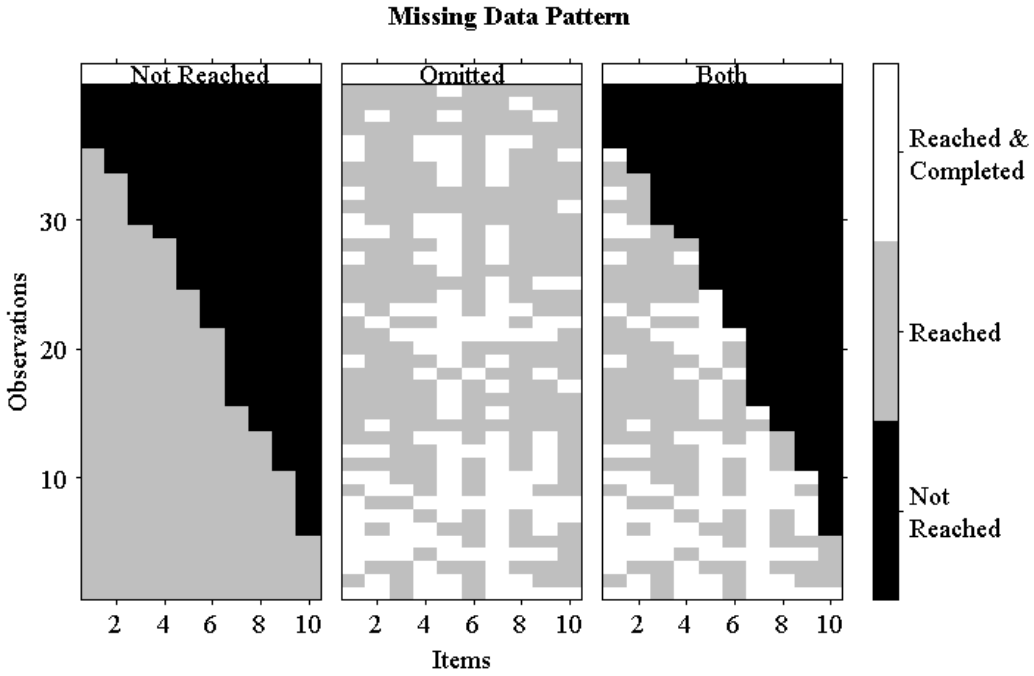

Figure 4.35: Missing data patterns due to not-reached items, omitted items, or both.

missing patterns have implications with respect to the appropriateness of missing data methods to handle item nonresponses. In the case of not-reached items, $\boldsymbol{D}$ can always be arranged to follow a perfect Guttman pattern. Such a pattern indicates particular dependencies between the missing indicator variables D_i. Let the index i indicate the position of the items in the test. If item i is the first item not reached by a test taker, then the probability to complete item Y_{i+1} is zero. Hence, $P(D_{i+1} = 1 \,|\, D_i = 0) = 0$. In contrast, $P(D_{i-1} = 1 \,|\, D_i = 0) = 1$. This is trivial since Y_{i-1} is always reached if item i is the first

not-reached item. Without further assumptions, this implies conditional stochastic independence $D_{i\neq k} \perp U \,|\, D_i$ as well as $D_{i\neq k} \perp (\boldsymbol{\xi}, \boldsymbol{\theta}) \,|\, D_i$ since $(\boldsymbol{\xi}, \boldsymbol{\theta}) = f(U)$. This violates the essential assumption of conditional stochastic independence $D_i \perp (\boldsymbol{D}_{-i}, \boldsymbol{Y}) \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$ in all MIRT models for nonignorable missing data discussed in this dissertation. Consequently, not-reached items should not be used as indicators in stochastic measurement models of latent response propensity. Only missing responses due to omissions can be appropriately handled by MIRT models for nonignorable missing data, since no deterministic relations between response indicator variables are implied in this case. The underlying conditional stochastic independence assumptions can potentially be met if the appropriate dimensionality of $\boldsymbol{\theta}$ is found and the correct model is specified.

Modelling not-reached items Conclusively, nonignorable missing responses due to not-reached items need to be taken into account in a different way. Glas and Pimentel (2008) proposed a special MIRT model for speeded tests, which typically suffer from substantial proportions of not reached items. This model is not considered in detail here. It should only be noted that the vector $\boldsymbol{D}$ is modeled by a sequential model (Tutz, 1997), which is closely related to the steps model for ordinal items (Verhelst, Glas, & De Vries, 1997). In both models it is assumed that the items consist of more than two ordered response categories and that each item is solved step by step. In the steps model , each response category is regarded as a Rasch-like item, where the item indicating response category h is only administered if $h-1$ was solved successfully. According to this idea, Glas and Pimentel adapted the matrix of indicator variables where only the first not-reached item is $D_i = 0$, all previous response indicators are $D_{j<i} = 1$, and all $D_{j>i}$ are treated as missing values. Hence, the matrix of response indicators contains missing data. Additionally, certain restrictions with respect to the thresholds in the sequential model are required. For more details of this model, see Glas and Pimentel (2008). The advantage of this approach is that the violation of local stochastic independence is taken into account. Unfortunately, the combination of sequential models and 1- or 2PLMs are hardly available in existing software.

In Section 4.5.4 a latent regression IRT model with $E[\boldsymbol{\xi} \,|\, f(\boldsymbol{D})]$ was proposed as an alternative to complex MIRT models. If the same set of items is applied to all test takers and missing data result exclusively from not-reached items, then the number of possible missing patterns $\boldsymbol{D} = \boldsymbol{d}$ is equal to $I + 1$, with I as the number of manifest items Y_i. That is, the number of responded items can range from zero to I. Since $\boldsymbol{D}$ always follows a perfect Guttman pattern (see Figure 4.35), all information of the missing pattern $\boldsymbol{D} = \boldsymbol{d}$ is

already given by the number of reached or not-reached items. Hence, the sum score S_D of the response indicators can be used as an appropriate function $f(\boldsymbol{D})$ in a latent regression model $E(\boldsymbol{\xi} \mid S_D)$. In the case of not-reached items, not only $c = f(\boldsymbol{D})$ but also $\boldsymbol{D} = f(S_D)$, implying conditional stochastic independence $\boldsymbol{D} \perp \boldsymbol{Y_{mis}} \mid (S_D, \boldsymbol{Y_{obs}})$. From Equation 4.151 follows that ML estimation is unbiased given no DIF exists in the measurement model of $\boldsymbol{\xi}$ depending on S_D.

If $\boldsymbol{\xi}$ is M-dimensional, then the regression $E(\boldsymbol{\xi} \mid S_D)$ consists of M univariate regressions $E[\xi_m \mid S_D]$, with $m = 1, \ldots, M$. In real applications, each of these regressions need to be correctly specified. Possible non-linear dependencies can be taken into account by polynomial regressions $E[\boldsymbol{\xi} \mid S_D] = b_0 + b_1 S_D + \ldots + b_k S_D^k$. As shown in Section 4.5.4, the LRM can easily be extended (see page 235). For example, further covariates represented by $\boldsymbol{Z}$ can be included in a multiple latent regression $E[\boldsymbol{\xi} \mid S_D, \boldsymbol{Z}]$. More complex item designs, such as balanced incomplete block designs can be taken into. Since each booklet consists of a different selection of items, the distribution of S_D as well as the stochastic relationship between $\boldsymbol{\xi}$ and S_D might vary across the test booklets. In this case, the booklet itself serves as a moderator variable. This can be taken into account by the inclusion of indicator variables $I_1, \ldots, I_H$ of the test booklets $1, \ldots, h, \ldots, H$ and interaction terms $_D \cdot I_1, \ldots, _D \cdot I_H$ in the latent regression. Alternatively, a multiple group model can be applied where the assigned booklets are used as the grouping variable. Under the assumption of measurement invariance, the item parameters are constrained to be equal across the groups. Additionally, group-specific regression $E_h(\boldsymbol{\xi} \mid S_D)$ are specified whose parameters are allowed to vary across groups to account for interaction effects between the assigned booklet and S_D with respect to $\boldsymbol{\xi}$. This model is less restrictive than the single group model with $E(\boldsymbol{\xi} \mid S_D, I_1, \ldots, I_H)$ since heterogeneous variances and covariances of the latent residuals ζ_m are allowed across groups.

The model proposed by Glas and Pimentel as well as the LRM and its extensions discussed here are limited to situations where item nonresponses result purely from not-reached items. However, in most real educational and psychological testings, missing data arise from both - intentional omissions of items and failing to complete the test due to time limits or a lack of motivation. Such situations result in complex missing data patterns such as the graphically represented one in the right graph of Figure 4.35. Here it is argued that missingness needs to be modeled differently depending on the reason of item nonresponse - omitted or not-reached items. A required joint model for omitted and not-reached items is proposed that accounts for respective peculiarities of both kinds of item nonresponses.

4.5.6.2 Developing a Joint Model of Omitted and Not-reached Items

In order to develop a joint model of omitted and not reached items, it is necessary to distinguish between $\boldsymbol{D}^{(O)}$ and $\boldsymbol{D}^{(N)}$. $\boldsymbol{D}^{(O)}$ is the I-dimensional response indicator variable with the elements $D_i^{(O)}$ which indicates whether item i is omitted or not, so that

$$D_i^{(O)} = \begin{cases} 1, \text{if } Y_i \text{ is } \textit{not} \text{ omitted} \\ 0, \text{if } Y_i \text{ is omitted.} \end{cases} \tag{4.160}$$

$\boldsymbol{D}^{(N)}$ is the I-dimensional response indicator variable with the elements $D_i^{(N)}$ which indicate whether item i is reached or not. That is,

$$D_i^{(N)} = \begin{cases} 1, \text{if } Y_i \text{ is reached} \\ 0, \text{if } Y_i \text{ is not reached.} \end{cases} \tag{4.161}$$

Note that neither $D_i^{(O)}$ nor $D_i^{(N)}$ indicate the observational status of Y_i that is given by D_i (see Equation 2.2). An item response is only observable if the item is reached *and* not omitted by the test taker. Hence, D_i is a function $f(D_i^{(O)}, D_i^{(N)})$ given by the following assignment rule

$$D_i = \begin{cases} 1, \text{if } D_i^{(O)} = 1 \text{ and } D_i^{(N)} = 1 \\ 0, \text{if } D_i^{(O)} = 0 \text{ and/or } D_i^{(N)} = 0. \end{cases} \tag{4.162}$$

Accordingly, the probability to observe an item response to item i is

$$P(D_i = 1) \quad = \quad P(D_i^{(O)} = 1 \cap D_i^{(N)} = 1), \tag{4.163}$$

and the counter-probability of a missing response is

$$P(D_i = 0) \quad = \quad P(D_i^{(O)} = 0 \cup D_i^{(N)} = 0). \tag{4.164}$$

In section 4.5.1 it was examined that $\boldsymbol{D}$ needs to be modeled jointly with $\boldsymbol{Y}$ if the missing data mechanism is NMAR. Instead of $\boldsymbol{D}$, the variables $\boldsymbol{D}^{(O)}$ and $\boldsymbol{D}^{(N)}$ are included in the model, which cover both the information about missingness and the reason of item nonresponses - omitted or not-reached. Hence, ML estimation is based on a the joint distribution $g(\boldsymbol{Y}, \boldsymbol{D}^{(O)}, \boldsymbol{D}^{(N)})$. An adequate model for omitted and not-reached items can be derived considering some peculiarities of the variables in the model.

Let there be a sample of N test takers answering the same set of items in the same order. That is, the single unit trial (see Equation 2.8) is repeated N times. The matrix of item responses $\boldsymbol{Y} = \boldsymbol{y}$ as well as $\boldsymbol{D}^{(O)} = \boldsymbol{d}^{(O)}$ and $\boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}$ are $N \times I$-matrices. The general ML estimator of the complete data is

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}, \boldsymbol{d}^{(O)}, \boldsymbol{d}^{(N)}; \boldsymbol{\iota}, \boldsymbol{\phi}) &\propto g(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{D}^{(O)} = \boldsymbol{d}^{(O)}, \boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}; \boldsymbol{\iota}, \boldsymbol{\phi}) \\ &\propto g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis}, \boldsymbol{D}^{(O)}_{obs} = \boldsymbol{d}^{(O)}_{obs}, \boldsymbol{D}^{(O)}_{mis} = \boldsymbol{d}^{(O)}_{mis}, \boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}; \boldsymbol{\iota}, \boldsymbol{\phi}) \end{aligned} \tag{4.165}$$

Due to factorization, this can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}, \boldsymbol{d}^{(O)}, \boldsymbol{d}^{(N)}; \boldsymbol{\iota}, \boldsymbol{\phi}) \propto\ & g(\boldsymbol{D}^{(O)}_{obs} = \boldsymbol{d}^{(O)}_{obs}, \boldsymbol{D}^{(O)}_{mis} = \boldsymbol{d}^{(O)}_{mis} \mid \boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}; \boldsymbol{\phi}) \\ & \cdot g(\boldsymbol{Y}_{obs} = \boldsymbol{y}_{obs}, \boldsymbol{Y}_{mis} = \boldsymbol{y}_{mis} \mid \boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}; \boldsymbol{\iota}, \boldsymbol{\phi}) \\ & \cdot g(\boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}; \boldsymbol{\phi}). \end{aligned} \tag{4.166}$$

The first conditional distribution refers to the model of the indicator vector $\boldsymbol{D}^{(O)}$ depending on $\boldsymbol{Y}$ and $\boldsymbol{D}^{(N)}$, the second conditional distribution refers to the target model of $\boldsymbol{Y}$ which is of substantial interest, and the last distribution refers to the distribution of the indicator vector $\boldsymbol{D}^{(N)}$. Interestingly, this model is kind of a mixture of a selection model and a pattern mixture model. However, Equation 4.166 serves only as a starting point in order to derive the final model for omitted and not-reached items. In the beginning of his section it was shown that the indicators $D_i^{(N)}$ of $\boldsymbol{D}^{(N)}$ should not be used in a probabilistic measurement model that indicates a latent response propensity. However, it could also be demonstrated that all information of $\boldsymbol{D}^{(N)}$ is already given by the sum score $S^{(N)}$, which is a function $f(\boldsymbol{D}^{(N)})$. Hence, $S^{(N)}$ can be used in a LRM. On the other hand, MIRT models for item nonresponses as derived in Section 4.5.3 are appropriate for omitted responses. Hence, omitted responses could be appropriately taken into account by a measurement model of a latent response propensity $\boldsymbol{\theta}$ based on $\boldsymbol{D}^{(O)}$. Hence, a joint model for nonignorable item nonresponses due to omitted and not-reached items is an MIRT model with two latent variables $\boldsymbol{\xi} = \xi_1, \ldots, \xi_M$ and $\boldsymbol{\theta} = \theta_1, \ldots, \theta_P$ indicated by $\boldsymbol{D}^{(O)}$ that includes a latent regression model with $S^{(N)}$ as the independent variable. Note that $\boldsymbol{\theta}$ represents the person?s tendency to respond to test items irrespective of whether the items are reached or not. Hence, missing data potentially occur not only in manifest variables Y_i but also in $D_i^{(O)}$ if not-reached items exist. The black squares in the right graph of Figure 4.35 indicate missing data in both $\boldsymbol{y}$ and $\boldsymbol{d}^{(O)}$. As in the case of $\boldsymbol{Y}$, it is important to know whether the missing data mechanism w.r.t. $\boldsymbol{D}^{(O)}$ is ignorable or not. It is likely that the number of not-reached items and the probability to omit a particular item are stochasti-

cally dependent. For example, let there be a timed achievement test. The tendency to omit items might be related to the probability to answer the items correct, implied by stochastic dependency between the latent ability and the latent response propensity. It is reasonable to assume that persons with lower ability levels need more time to process single items, leading to a lower number of items that can be reached within the time limit. In this case, the proportion of omitted items and the number of not-reached items are positively correlated. Nevertheless, there might be applications where the omission rate increases with decreasing numbers of not-reached items. For example, in low-stakes assessment, persons with lower proficiency levels might show higher rates of omissions. If test takers quickly decide to omit the items, then they might reach the end of the test even faster than more proficient test takers who try to answer items carefully. In this case, the correlation of the proportion of not-reached items with the latent ability as well as with the omission rate would be negative. The tendency to omit items, however, would still be positively correlated with the persons' latent ability. Other conditions might be plausible as well. In any case, the missing data mechanism w.r.t. $\boldsymbol{D}^{(\boldsymbol{O})}$ should be expected to be dependent on the number of not-reached items, suggesting nonignorable missing data in $\boldsymbol{d}^{(\boldsymbol{O})}$. The missing data mechanisms w.r.t. $\boldsymbol{D}^{(\boldsymbol{O})}$ can be equivalently defined as in the case of $\boldsymbol{Y}$. The only difference is that $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ needs to be taken into account. Considering the joint distribution of $(\boldsymbol{D}^{(\boldsymbol{O})}_{obs}, \boldsymbol{D}^{(\boldsymbol{O})}_{mis}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, S^{(N)})$, the missing data mechanism w.r.t. $\boldsymbol{D}^{(\boldsymbol{O})}$ is

(a) MCAR if $\boldsymbol{D}^{(\boldsymbol{O})} \perp (S^{(N)}, \boldsymbol{Y})$,

(b) MAR if $\boldsymbol{D}^{(\boldsymbol{O})}_{mis} \perp S^{(N)} \,|\, (\boldsymbol{D}^{(\boldsymbol{O})}_{obs}, \boldsymbol{Y}_{obs})$, and

(c) NMAR if $\boldsymbol{D}^{(\boldsymbol{O})}_{mis} \not\perp S^{(N)} \,|\, (\boldsymbol{D}^{(\boldsymbol{O})}_{obs}, \boldsymbol{Y}_{obs})$.

If additional observed covariates are included (i. e. in a background model), then the definitions are quite close; The missing data mechanism w.r.t. $\boldsymbol{D}^{(\boldsymbol{O})}$ is

(a) MCAR if $\boldsymbol{D}^{(\boldsymbol{O})} \perp (S^{(N)}, \boldsymbol{Y}, \boldsymbol{Z})$,

(b) MAR if $\boldsymbol{D}^{(\boldsymbol{O})}_{mis} \perp S^{(N)} \,|\, (\boldsymbol{D}^{(\boldsymbol{O})}_{obs}, \boldsymbol{Z}, \boldsymbol{Y}_{obs})$, and

(c) NMAR if $\boldsymbol{D}^{(\boldsymbol{O})}_{mis} \not\perp S^{(N)} \,|\, (\boldsymbol{D}^{(\boldsymbol{O})}_{obs}, \boldsymbol{Y}_{obs}, \boldsymbol{Z})$.

For reasons of clarity, however, covariates are not included in the following derivations. In a joint model for missing responses due to omitted and not-reached items, the missingness

in $\boldsymbol{Y}$ and $\boldsymbol{D}^{(\boldsymbol{O})}$ needs to be taken into account. The LRM with $E(\boldsymbol{\xi} \,|\, S^{(N)})$ was found to be appropriate to account for item-nonresponses in $\boldsymbol{Y}$ due to not-reached items. Based on the same rationale, the LRM with $E(\boldsymbol{\theta} \,|\, S^{(N)})$ accounts for missing data in $\boldsymbol{D}^{(\boldsymbol{O})}$ given certain assumptions hold true. In particular, it is assumed that there is no DIF with respect to the manifest variables Y_i and $D_i^{(O)}$ depending on $S^{(N)}$ and other manifest variables in the model. Furthermore, local stochastic independence of all Y_i and D_i, with $i = 1, \ldots, I$, is assumed. Formally, that means that

$$Y_i \perp (\boldsymbol{Y}_{-i}, \boldsymbol{D}, \boldsymbol{\theta}, S^{(N)}) \quad | \quad \boldsymbol{\xi} \qquad \forall i = 1, \ldots, I, \tag{4.167}$$

and

$$D_i^{(O)} \perp (\boldsymbol{D}_{-\boldsymbol{i}}^{(\boldsymbol{O})}, \boldsymbol{Y}, S^{(N)}) \quad | \quad \boldsymbol{\theta} \qquad \forall i = 1, \ldots, I. \tag{4.168}$$

The MML estimator of the complete data is then

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}, \boldsymbol{d}^{(\boldsymbol{O})}, \boldsymbol{s}^{(\boldsymbol{N})}; \boldsymbol{\iota}, \boldsymbol{\phi}) \quad \propto \quad & \prod_{n=1}^{N} \int_{\mathbb{R}^{M \times P}} \Big[g(\boldsymbol{D}_{\boldsymbol{n;obs}}^{(\boldsymbol{O})} = \boldsymbol{d}_{\boldsymbol{n;obs}}^{(\boldsymbol{O})}, \boldsymbol{D}_{\boldsymbol{n;mis}}^{(\boldsymbol{O})} = \boldsymbol{d}_{\boldsymbol{n;mis}}^{(\boldsymbol{O})} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}) \\ & \cdot g(\boldsymbol{Y}_{\boldsymbol{n;obs}} = \boldsymbol{y}_{\boldsymbol{n;obs}}, \boldsymbol{Y}_{\boldsymbol{n;mis}} = \boldsymbol{y}_{\boldsymbol{n;mis}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota}) g(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, S_n^{(N)} = s_n^{(N)}; \boldsymbol{\phi}) \Big] d(\boldsymbol{\xi}, \boldsymbol{\theta}). \end{aligned} \tag{4.169}$$

Compared to the Equations 4.165 and 4.166, the distribution $g(\boldsymbol{D}^{(N)} = \boldsymbol{d}^{(N)}; \boldsymbol{\phi})$ has been skipped from the MML estimatior since $\boldsymbol{D}^{(N)}$ is replaced by $S^{(N)}$, which is used as a purely exogenous variable in the two latent regression models $E(\boldsymbol{\xi} \,|\, S^{(N)})$ and $E(\boldsymbol{\theta} \,|\, S^{(N)})$. As in common regression models, the distribution of the independent variables are not required to be modeled. Following the ideas of the MIRT models for omitted responses, the conditional distributions and $g(\boldsymbol{D}_{\boldsymbol{obs}}^{(\boldsymbol{O})} = \boldsymbol{d}_{\boldsymbol{obs}}^{(\boldsymbol{O})}, \boldsymbol{D}_{\boldsymbol{mis}}^{(\boldsymbol{O})} = \boldsymbol{d}_{\boldsymbol{mis}}^{(\boldsymbol{O})} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi})$ are replaced by I model equations for the response indicators $D_i^{(O)}$. Here, the one- and the two-parameter models are proposed, so that

$$P(D_i^{(O)} = 1 \,|\, \boldsymbol{\theta}; \boldsymbol{\phi}) \quad = \quad \frac{\exp(\boldsymbol{\gamma}_{\boldsymbol{i;\theta}} \boldsymbol{\theta} - \gamma_{i0})}{1 + \exp(\boldsymbol{\gamma}_{\boldsymbol{i;\theta}} \boldsymbol{\theta} - \gamma_{i0})}. \tag{4.170}$$

Alternatively, one of the within-item multidimensional 1- or 2PL models as proposed in Section 4.5.3.3 can be used. In this case, the assumption given by Equation 4.168 is modified so that conditional stochastic independence $D_i^{(O)} \perp (\boldsymbol{D}_{-\boldsymbol{i}}^{(\boldsymbol{O})}, \boldsymbol{Y}, S^{(N)}) \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$ is

assumed for all $i = 1, \ldots, I$

$$P(D_i^{(O)} = 1 \,|\, \boldsymbol{\theta}^*, \boldsymbol{\xi}; \boldsymbol{\phi}) \quad = \quad \frac{\exp(\boldsymbol{\gamma_i}(\boldsymbol{\xi}, \boldsymbol{\theta}^*)^T - \gamma_{i0})}{1 + \exp(\boldsymbol{\gamma_i}(\boldsymbol{\xi}, \boldsymbol{\theta}^*)^T - \gamma_{i0})}. \tag{4.171}$$

The conditional distribution $g(\boldsymbol{Y_{obs}} = \boldsymbol{y_{obs}}, \boldsymbol{Y_{mis}} = \boldsymbol{y_{mis}} \,|\, \boldsymbol{\xi}; \boldsymbol{\iota})$ refers to the target model, that is, the measurement model of the latent ability $\boldsymbol{\xi}$. Accordingly, the respective model equations of the IRT measurement model used for the items Y_i, such as the uni- or multidimensional one- or two-parameter models, are used. The observed data likelihood results from the complete data likelihood by integrating over all possible values of $(\boldsymbol{Y_{mis}}, \boldsymbol{D_{mis}^{(O)}})$ that are consistent with the observed missing data pattern $\boldsymbol{D} = \boldsymbol{d}$ (see Section 4.5.1). Inserting the model equations of the items Y_i and the indicators $D_i^{(O)}$, the observed data likelihood $\mathcal{L}(\boldsymbol{y_{obs}}, \boldsymbol{d_{obs}^{(O)}}, \boldsymbol{s^{(N)}}; \boldsymbol{\iota}, \boldsymbol{\phi})$ of the joint model for omitted and not-reached items can be written as

$$\begin{aligned} L(\boldsymbol{y_{obs}}, \boldsymbol{d_{obs}^{(O)}}, \boldsymbol{s^{(N)}}; \boldsymbol{\iota}, \boldsymbol{\phi}) \quad = \quad & \prod_{n=1}^{N} \int_{\mathbb{R}^{P \times M}} \Big\{ \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \,|\, \boldsymbol{\xi}, \boldsymbol{\iota})^{d_{ni}^{(O)} \cdot d_{ni}^{(N)}} P(D_{ni}^{(O)} = d_{ni}^{(O)} \,|\, \boldsymbol{\theta}; \boldsymbol{\phi_1})^{d_{ni}^{(N)}} \\ & g(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, S_n^{(N)} = s_n^{(N)}; \boldsymbol{\phi_2}) \Big\} d\boldsymbol{\xi} d\boldsymbol{\theta}. \end{aligned} \tag{4.172}$$

Because of the exponents $d_{ni}^{(O)} \cdot d_{ni}^{(N)}$ and $d_{ni}^{(N)}$, only the observed responses of Y_i and $D_i^{(O)}$ remain in the observed data likelihood function. Note that the parameter vector $\boldsymbol{\phi}$ has been divided into two parts $\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$. $\boldsymbol{\phi_1}$ consists of the parameters of the model of $\boldsymbol{D^{(O)}}$, such as $\boldsymbol{\gamma}$ and the thresholds γ_{i0}. The parameters in $\boldsymbol{\phi_2}$ consist of the parameters of the two latent regressions $E(\boldsymbol{\xi} \,|\, S^{(N)})$ and $E(\boldsymbol{\theta} \,|\, S^{(N)})$, which are the regression coefficients, intercepts and residual variances $Var(\zeta_{\xi_m})$ and $Var(\zeta_{\theta_l})$, and covariances $Cov(\zeta_{\xi_m}, \zeta_{\xi_{j \neq m}})$,$Cov(\zeta_{\theta_l}, \zeta_{\theta_{k \neq l}})$ and $Cov(\zeta_{\xi_m}, \zeta_{\theta_l})$. ζ_{ξ_m} is the residual of the latent regression $E(\xi_m \,|\, S^{(N)})$, and ζ_{θ_l} is the residual of the latent regression $E(\theta_l \,|\, S^{(N)})$. The basic idea is that the conditional distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, S_n^{(N)} = s_n^{(N)}; \boldsymbol{\phi_2})$ is modeled by a multivariate distribution of the latent variables with the expected value $E[(\boldsymbol{\xi}, \boldsymbol{\theta}) \,|\, S^{(N)}]$. It is assumed that $(\boldsymbol{\xi}, \boldsymbol{\theta})$ is conditionally multivariate normal given $S^{(N)}$, so that the latent residual $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \boldsymbol{\theta})^T - E[(\boldsymbol{\xi}, \boldsymbol{\theta})^T \,|\, S^{(N)}]$ is normally distributed with $\boldsymbol{\zeta} \sim N(\mathbf{0}, \Sigma_\zeta)$. Σ_ζ is the variance-covariance matrix of $\boldsymbol{\zeta}$. In this form, the joint model for omitted and not-reached items can be estimated in available software packages that allow for estimating multidimensional IRT models with latent regressions, such as M*plus* (Muthén & Muthén, 1998 - 2010). Alternatively, other distributions than the multivariate normal distribution of $\boldsymbol{\zeta}$ could be chosen when MML estimation is used.

Example For illustration, the hypothetical example of the 2PL-BMIRT model depicted in Figure 4.23 is extended here to account for missing data due to omitted items and not-reached items. The resulting model is shown in Figure 4.36. Comparing the two models (cf. Figures 4.23 and 4.36) highlights the differences between the MIRT model for omitted items and the joint model for omitted and not-reached items. In the latter, the indicators $D_i^{(O)}$ instead of D_i constitute the measurement model of the latent response propensity $\boldsymbol{\theta}$. Furthermore, a latent regression model has been added with the number of reached items $S^{(N)}$ as predictor. Four linear[27] regressions $E(\xi_1 \mid S^{(N)})$, $E(\xi_2 \mid S^{(N)})$, $E(\theta_1 \mid S^{(N)})$,

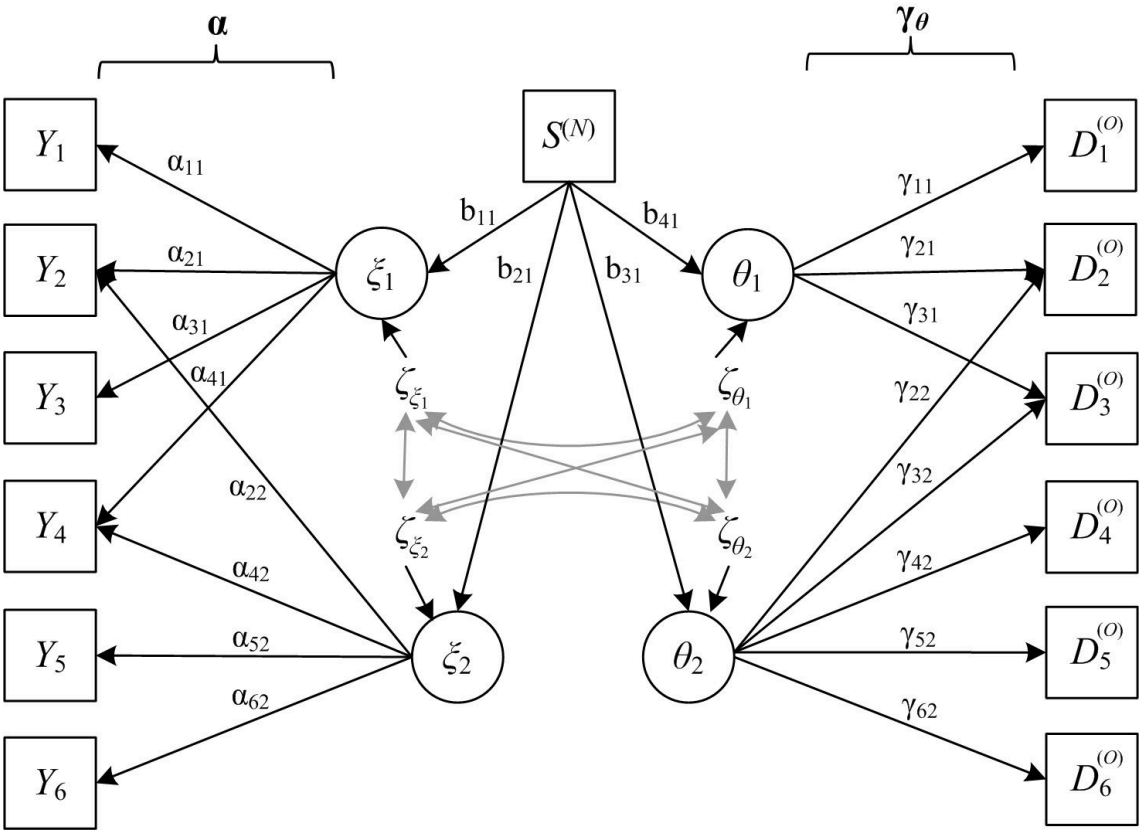

Figure 4.36: Graphical representation of the joint MIRT model for omitted and not-reached items.

and $E(\theta_2 \mid S^{(N)})$ are included. It is common in SEM and latent trait models to combine these regressions to multivariate regression $E[(\boldsymbol{\xi}, \boldsymbol{\theta})^T \mid S^{(N)}]$, with $(\boldsymbol{\xi}, \boldsymbol{\theta}) = (\xi_1, \xi_2, \theta_1, \theta_2)$.

[27] The LRM can easily be extended to the case of non-linear regressions by the inclusion of polynomial functions of $S^{(N)}$ as predictors in the LRM.

Accordingly, the vector of latent variables can be written as

$$(\boldsymbol{\xi}, \boldsymbol{\theta})^T = E[(\boldsymbol{\xi}, \boldsymbol{\theta})^T \,|\, S^{(N)}] + \boldsymbol{\zeta}. \tag{4.173}$$

In the case of linear regressions, that is,

$$(\boldsymbol{\xi}, \boldsymbol{\theta})^T = \mathbf{b_0} + \mathbf{b_1} S^{(N)} + \boldsymbol{\zeta} \tag{4.174}$$

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} b_{10} \\ b_{20} \\ b_{30} \\ b_{40} \end{pmatrix} + \begin{pmatrix} b_{11} \\ b_{21} \\ b_{31} \\ b_{41} \end{pmatrix} S^{(N)} + \begin{pmatrix} \zeta_{\xi_1} \\ \zeta_{\xi_2} \\ \zeta_{\theta_1} \\ \zeta_{\theta_2} \end{pmatrix}. \tag{4.175}$$

The non-diagonal elements in the covariance matrix Σ_ζ are equal to the conditional covariances $Cov(\xi_m, \xi_{j\neq m} \,|\, S^{(N)})$, $Cov(\theta_l, \theta_{k\neq l} \,|\, S^{(N)})$, and $Cov(\xi_m, \theta_l \,|\, S^{(N)})$, which are depicted as grey lines in Figure 4.36. If the model is correctly specified, and the model assumptions hold true, then the covariances $Cov(\xi_m, \theta_l \,|\, S^{(N)})$ can be used to study the conditional stochastic independence between the tendency to omit items and the latent ability given the number of not-reached items which are implied by the stochastic dependencies between latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. The regression coefficients of $E(\boldsymbol{\xi} \,|\, S^{(N)})$ are informative with respect to the regressive dependency of the latent ability on item nonresponses due to not-reached items. Hence, the parameters of the regressions as well as the covariance structure of the latent residuals are informative about the missing data mechanism. If all conditional covariances $Cov(\zeta_{\theta_l}, \zeta_{\xi_m} \,|\, S^{(N)}) = 0$ and the assumptions 4.167 and 4.168 are valid, then conditional regressive independence $Y_i \perp \boldsymbol{D^{(O)}} \,|\, S^{(N)}$ is implied for all variables Y_i[28]. In this case, missing data resulting from omissions of items are not required to be modeled jointly with $\boldsymbol{Y}$ if the regression $E(\boldsymbol{\xi} \,|\, S^{(N)})$ is included. The measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D^{(O)}}$ can be left out, which may simplify the model considerably. Caution is required in order to decide whether $S^{(N)}$ can be left out of the model, since regressive independence between $\boldsymbol{\xi}$ and $S^{(N)}$ is not sufficient. Recall that the missing data mechanism w.r.t. $\boldsymbol{D^{(O)}}$ is potentially nonignorable regardless of the stochastic relationship between $\boldsymbol{\xi}$ and $S^{(N)}$. This is the case if the number of reached items and the tendency for omissions indicated by $\boldsymbol{D^{(O)}}$ are stochastically dependent. The missing data mechanism w.r.t. $\boldsymbol{D^{(O)}}$ is then potentially MAR or even NMAR. $S^{(N)}$ needs to be included in the latent regression $E(\boldsymbol{\theta} \,|\, S^{(N)})$ to ensure unbiased parameter estimation in the model of $\boldsymbol{D^{(O)}}$,

[28] In the case of dichotomous variables, this conditional regressive independence implies conditional stochastic independence $\boldsymbol{Y} \perp \boldsymbol{D^{(O)}} \,|\, S^{(N)}$.

which is essential to account for nonignorable missing response w.r.t. $\boldsymbol{Y}$ due to omissions. Thus, given the missing data mechanism w.r.t. $\boldsymbol{Y}$ is NMAR due to omitted responses, $S^{(N)}$ can be left out only if two conditions hold true: (a) $\boldsymbol{Y}_{mis} \perp S^{(N)} \,|\, (\boldsymbol{Y}_{obs}, \boldsymbol{D}_{obs})$ and (b) $\boldsymbol{D}^{(O)}_{obs} \perp S^{(N)} \,|\, (\boldsymbol{Y}_{obs}, \boldsymbol{D}_{obs})$. Unfortunately, these conditions cannot be tested in application. It is recommended to skip the LRM with $S^{(N)}$ only if all regression coefficients of the multivariate regression $E[(\boldsymbol{\xi}, \boldsymbol{\theta})^T \,|\, S^{(N)}]$ are zero. Even though this assumption is stronger, it is testable in application.

Extensions of the model As in any regression model, valid inference depends on the correct specification of the model. Nonlinear regressions $E(\xi_m \,|\, S^{(N)})$ and $E(\theta_m \,|\, S^{(N)})$ using polynomial regressions might be required to model the stochastic dependencies appropriately. Alternatively, a saturated model using the indicator variables for the number of reached items could be used instead of $S^{(N)}$. However, this might result in unnecessarily complex models if the number of items becomes large. The model can also be extended by inclusion of additional covariates Z_j in the background model so that $E(\xi_m \,|\, S^{(N)}, \boldsymbol{Z})$ and $E(\theta_m \,|\, S^{(N)}, \boldsymbol{Z})$. As previously explained, this also allows to account for not-reached items in more complex test designs. For example, in balanced incomplete block designs, indicator variables for the different test booklets could be used. If the booklet moderates the stochastic dependency between the latent variables in the model - here $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ - then interaction terms with the indicators of the booklets and the number of not-reached items can be included in the latent regression model. Alternatively, the joint model for omitted and not-reached items can be combined with a multiple group model with the assigned test booklet as the grouping variable. Under the assumption of measurement invariance, the item parameters are constrained to be equal across the groups. The parameters of the regressions $E(\boldsymbol{\xi} \,|\, S^{(N)})$ and $E(\boldsymbol{\theta} \,|\, S^{(N)})$, however, can be different across groups to allow for interaction between the assigned booklet and $S^{(N)}$ with respect to the latent variable.

Practical recommendations Many other variants, combinations, and extensions of the models are thinkable, which were not discussed here. It should only be noted that a considerable flexibility exists in combining different models such as MIRT models, latent regression models, and multiple group models. In application, the aim is to find the most parsimonious but sufficient model that accounts for item nonresponses. This might become an increasingly challenging task in complex measurement models and complex test designs. In real applications it is strongly recommended to approach the final model step by step starting with exploring $\boldsymbol{D}$ and $\boldsymbol{D}^{(O)}$ respectively and finding an appropriate

model for the response indicator vector. The findings should be taken into account in the joint model of $(\boldsymbol{D}, \boldsymbol{Y})$ or $(\boldsymbol{D}^{(O)}, \boldsymbol{D}^{(N)}, \boldsymbol{Y})$ respectively. If the model becomes too complex, then the model complexity can be reduced. For example, instead of modeling $(\boldsymbol{D}^{(O)}, \boldsymbol{Y})$ jointly by a measurement model with two latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ and the latent regression model including $S^{(N)}$, person parameter estimates $\hat{\boldsymbol{\theta}}$ can be estimated in a first step using only the measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D}^{(O)}$ including the latent regression $E(\boldsymbol{\theta} \mid S^{(N)})$. In a second step, only the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ is estimated including the latent regression $E(\boldsymbol{\xi} \mid \hat{\boldsymbol{\theta}}, S^{(N)})$. In this case, $\boldsymbol{D}^{(O)}$ can be left out in the second step. This may substantially reduce the number of manifest and latent variables and the model complexity, respectively.

4.6 Discussion

The results of Chapter 3 clearly stressed the importance of appropriate methods to handle item nonresponses in IRT measurement models. In this chapter, existing ad-hoc and model based methods for item nonresponses have been studied in detail. These methods have been considered in light of missing data theory and well-known missing data methods, such as imputation methods and ML estimators for missing data. These methods are well-developed and are regarded as state-of-the-art (Schafer & Graham, 2002). It could be shown that ad-hoc methods, such as incorrect answer substitution (IAS) and partially correct scoring (PCS) of item nonresponses, can be regarded as deterministic single imputation methods with very strong assumptions not tenable in light of modern missing data theory. Although especially IAS has often been criticized (Lord, 1974, 1983a; Rose et al., 2010), it is still commonly used even in prestigious international large scale assessment, such as PISA 2006 (Organisation for Economic Co-operation and Development, 2009b, 2009a). For this reason, IAS was once more addressed here in this thesis. PCS was initially introduced as an alternative to IAS (Lord, 1974). However, IAS and PCS are very similar. In fact, IAS can formally be seen as a special case of PCS. To consider an item nonresponse as an additional response category in a nominal response model (NRM Moustaki & O'Muircheartaigh, 2000) is a more recent alternative approach also critically examined here. In a number of papers, multidimensional IRT models for nonignorable missing data have been discussed in the last years (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). These models are of major interest, since strong evidence exists that item nonresponses are related to the latent ability intended to be measured by

the test. In this case, the missing data mechanism w.r.t. $\boldsymbol{Y}$ is NMAR. Apart from educational and psychological measurement, selection models (SLM) and pattern mixture models (PMM) have been developed to account for nonignorable missing data. In this work it was shown that MIRT models for missing responses that are NMAR can be derived from both classes of models under certain assumptions. In multidimensional IRT (MIRT) models, a distinction is made between within- and between-item multidimensionality. This terminology is also used to classify MIRT models for item nonresponses. Between-item (B-MIRT) and within-item multidimensional IRT (B-MIRT) models were typically considered to be equivalent. As shown here, that is not per se true. Considering the issue of model equivalence, different 1PL- and 2PL-W-MIRT models were theoretically derived. The meanings of the latent variables in the different models were compared. With latent regression models (LRM) and multiple group (MG) IRT models resting on functions $f(\boldsymbol{D})$, more parsimonious and computationally less demanding approaches have been discussed. These approaches were introduced by Rose et al. (2010). Here the underlying rational is given in detail. MIRT- and MG-IRT models as well as LRMs are strongly related. It was emphasized that the underlying dimensional structure of $\boldsymbol{D}$ should be examined carefully, even if LRMs or MG-IRT models are used in order to find adequate functions $f(\boldsymbol{D})$ used as independent variables in the LRM or as grouping variables in MG-IRT model.

In MIRT models, local stochastic independence of the manifest variables Y_i and D_i is assumed. The assumption of local stochastic independence $D_i \perp D_j | \boldsymbol{\theta}$ is only reasonable for item nonresponses due to omitted items but necessarily violated in the case of not-reached items. Hence, MIRT models are only appropriate to account for omissions. Missing responses caused by not-reached items can be properly handled using LRMs. Since item nonresponses typically result from both, omitted and not-reached items, a joint model has been introduced to account for these two reasons of missing responses. The main results and conclusions of this chapter are summarized next.

4.6.1 Ad-hoc Methods for Item Nonresponses

Ad-hoc techniques to handle item nonresponses are still commonly used in psychological and educational measurement. IAS and PCS are such methods. Unfortunately, these methods seem to be very plausible, which may explain their popularity. Nevertheless, especially IAS was repeatedly criticized (e. g. Culbertson, 2011, April; Lord, 1974, 1983a; Rose et al., 2010). For that reason, both IAS and PCS were once more scrutinized analytically and by means of simulated data.

Incorrect answer substitution (IAS) In Chapter 3 it could be demonstrated that IAS is implicitly applied when the sum score is used as a test score. The alarming results with respect to bias, validity, and test fairness of the sum score under any missing data mechanism suggest that IAS similarly affects IRT-based item and person parameters. Following Huisman (2000), IAS can be seen as a naive imputation technique. It is well known that the unbiasedness of sample-based parameter estimates depends fundamentally on the validity of the imputation model (e. g. Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002). The imputation model underlying IAS is quite simple; each item nonresponse is recoded to an incorrect answer. This naive imputation model is correct if the assumption $P(Y_i = 1 \,|\, D_i = 0) = 0$ holds true. However, a closer examination reveals that this assumption is hardly tenable in real applications. Test takers have different strategies to answer a test. Persons may change the order in which they respond to the items. Some start with answering easier items and processing more demanding items afterwards. Especially in timed tests, this can result in missing data. However, the persons have a probability $P(Y_i = 1 \,|\, D_i = 0) = P(Y_i = 1 \,|\, \boldsymbol{\xi})$ to solve the item given their proficiency levels. Furthermore, in low-stakes assessments, the willingness for exertion is, on average, pretty low. Items that are demanding are more likely to beleft out. However, this does not necessarily imply that the test takers would have had no chance to answer the item correctly if they had tried to answer it seriously. Proponents of IAS argue that persons decide to omit items if they feel that they cannot solve an item. Even in this case, IAS is highly questionable, because IAS implicitly assumes that test takers process items completely and can perfectly judge whether the result is correct or not. That is, if they „know“ that the result is incorrect, then they decide to skip the item. This seems unrealistic. Persons may have a feeling about their performance in the test and on particular items but rarely have a perfect knowledge about the correctness of their item responses. Apart from these more informal considerations, IAS is even more dubious considering the formal implications. The implicit assumption of conditional independence $P(Y_i = 1 \,|\, D_i = 0) = 0$ under IAS conflicts with the IRT model that is aimed to be estimated. This was shown in detail by analytical means. The IRT model equation describes the probability $P(Y_i = 1 \,|\, \boldsymbol{\xi})$ regardless of the observational status of Y_i. Given that no DIF exist with respect to D_i, it follows that $P(Y_i = 1 \,|\, \boldsymbol{\xi}, D_i = 1) = P(Y_i = 1 \,|\, \boldsymbol{\xi}, D_i = 0) = P(Y_i = 1 \,|\, \boldsymbol{\xi})$. This is incompatible with $Y_i \perp \boldsymbol{\xi} \,|\, D_i = 0$ implied by $P(Y_i = 1 \,|\, D_i = 0) = 0$ under IAS. Finally, it could be demonstrated that standard ML estimation of item and person parameters applied to filled-in data sets does not account for the implications of IAS. Conclusively, item and person parameter estimates are biased. It could be shown that the functional form of ICCs

under IAS strongly depends on the missing data mechanism w.r.t. Y_i when IAS is used. If the nonresponse mechanism is MCAR, then an upper and a lower asymptote are implied, so that none of the one- to three-parametric IRT models are appropriate.

In the considerations of the sum score, it was already shown that IAS can also be seen as a replacement of items Y_i by the product variables $Y_i^* = Y_i \cdot D_i$. If the assumption $P(Y_i = 1 \mid D_i = 0) = 0$ does not hold true, the variables Y_i and Y_i^* as well as their distributions are different. Strictly speaking, a different random experiment (see Section 2.1) is considered if IAS is used that includes an additional step: When Y_i is not observable, assign D_i the value 0 *and* set $Y_i = 0$. To change the random experiment means to consider a different measurement model based on different manifest variables. Since the latent variable $\boldsymbol{\xi}$ is constructed in a measurement model rather than simply measured, the meaning of the latent variable may change. Indeed, it could be shown that the variables Y_i^* confound information about test performance and willingness/ability to respond to the items. Accordingly, the latent variable $\boldsymbol{\xi}^*$ constructed in a measurement model based on Y_i^* is a mixture of the latent proficiency $\boldsymbol{\xi}$ intended to be measured, and the latent response propensity $\boldsymbol{\theta}$. The findings were illustrated by simulated data and highlight that missing responses as well as the handling of them are potentially a threat of validity. Due to the results of this thesis, IAS cannot be recommended in any application because of unrealistic assumptions and theoretical inconsistencies with stochastic IRT measurement models.

Partially correct scoring (PCS) To overcome the implausible assumptions of IAS, Lord (Lord, 1974) proposed a scoring of item nonresponses as partially correct. He assumed that each person that did not process an item i has a positive probability $P(Y_i = 1 \mid D_i = 0) = c$ to answer this item. Unfortunately, PCS suffers from similar inconsistencies as IAS. In fact, here it could formally be shown that IAS is a special case of PCS with $c = 0$. PCS also assumes $Y_i \perp \boldsymbol{\xi} \mid D_i = 0$ implicitly. This is implied by the two underlying assumptions explicitly made by Lord: (a) a nonresponse occurs only if a test taker does not know the answer and (b) if a test taker would have answered the missing item, he or she would have guessed. This reflects Lord's assumption that a person who did not answer item i is totally undecided which response category is the correct one. Consequently, the potential response is purely at random, so that $c = 1/k$ is a constant for all persons regardless of their proficiency level. k is the number of response categories. Considering the likelihood function, it reveals that PCS is equivalent to replace nonresponses by c. For that reason, PCS is also a naive imputation method. In his original paper, Lord proofed that PCS scoring is equivalent to impute missing responses by random draws of Y_i with

the probability c for each response category if $N \to \infty$. However, such random draws introduce noise into data and may lower stochastic dependencies between variables. Accordingly, it was hypothesized that item discriminations are increasingly underestimated with higher proportions of item nonresponses if PCS is used. This could be confirmed by simulated data. The item discriminations were negatively biased and tend toward zero with rising proportions of missing responses. The latent variable $\hat{\xi}_{PCS}$ is also constructed differently. Under the conditions considered here, $\hat{\xi}_{PCS}$ is nonlinear function of both the latent ability ξ of interest and the latent response propensity θ. Finally, it was shown that standard ML estimators of IRT parameters used in combination with PCS have improper characteristics. As in the case of IAS, standard ML estimation methods do not account for the implications of PCS.

Therefore, PCS as well as IAS have been proved to be inappropriate in almost all real applications. Due to theoretical inconsistencies as well as implausible assumptions, the use of IAS and PCS is generally not recommended.

Nominal Response Model To consider an item nonresponse as an additional response category is an approach that ranges between data augmentation methods and model based methods. With data augmentation methods it shares the property that the data matrix used for model parameter estimation does not contain missing data anymore. The approach can also be seen as model based method, since an explicit stochastic model for item nonresponses is included. This approach seems to be promising since missing data are modeled by an additional response category and allows for stochastic dependencies between D_i and $\boldsymbol{\xi}$. In fact, by means of simulated data, it could be shown that the item parameters of the measurement model of a unidimensional latent variable ξ can be estimated unbiasedly when the values of ξ are known. This is unrealistic in almost all real application. If both item parameters and individuals values of the latent ability are unknown and aimed to be estimated, then similar problems arise as in the case of IAS and PCS. The reason is that the manifest items Y_i are replaced by a new random variable - here denoted by R_i. If Y_i is dichotomous, then the variables R_i used in NRM have three response categories: $R_i = 0$ indicates a wrong response, $R_i = 1$ indicates a correct answer, and $R_i = 2$ when the item response is missing. Therefore, R_i is a function $f(Y_i, D_i)$. As in the case of IAS, the manifest variables in the measurement confound two types of information: the performance in the test given by correct and incorrect responses and the willingness or the ability to provide any response. Recall that the latent variable is not simply measured but constructed on the basis of the manifest variables that refer to a particular random experiment. In the

NRM, the constructed latent variable is also a kind of linear combination of ξ and the latent response propensity θ. This reflects the fusion of two different variables Y_i and D_i into R_i, the variables in the measurement model of ξ_{NRM}. It could be shown analytically and empirically that the NRM will be useful in one particular case. That is, if the correlation of the latent response propensity θ and the latent variable ξ is equal to one. That means that the tendency to respond to item i depends solely on the latent ability. However, in this case the use of the NRM is without any added value, since the response indicators D_i could also be used as additional items indicating ξ directly. If the latent response propensity θ and the latent variable ξ are not linear functions of each other, then the NRM will fail to estimate the correct item and person parameters. Note that the term „biased“ parameter estimates might be misleading here. The problem is not that the sample based estimates suffer from inconsistency. Rather, the *true* model parameters in the IAS and NRM are different from those in the measurement model of ξ based on $\boldsymbol{Y}$, which are actually of interest and aimed to be estimated in application. The measurement models using $\boldsymbol{Y}$, $\boldsymbol{Y}^*$, or $\boldsymbol{R}$ consist of different manifest variables with different distributions, leading to differently constructed latent variables. Thus, strictly speaking, the model parameters and their sample-based estimates become incomparable. From these findings it is concluded that the use of the NRM based on R_i is also questionable in many real applications. The implicit assumption of a perfect correlation between the tendency to respond to the items and the latent ability is very strong. Less restrictive model-based approaches should be preferred.

4.6.2 Model Based Approaches

In the literature of missing data methods, it is generally distinguished between traditional methods (e.g. listwise and pairwise deletion), imputation methods (e.g. regression imputation, multiple imputation), and model based methods (e.g. FIML) (Lüdtke et al., 2007). These classifications can also be used for approaches to handle item nonresponses in measurement models. As previously noted, IAS and PCA are naive imputation methods. With multiple imputation (MI) elaborated data augmentation methods have been developed, which have proved to be useful in IRT measurement models as well even if the proportion of missing data is large (Van Buuren, 2010). However, MI requires that the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR[29]. Standard ML estimation methods, such as JML and MML, can be regarded as an FIML estimator, since each observed item re-

[29]The missing data mechanism w.r.t. $\boldsymbol{Y}$ can be MAR given $\boldsymbol{Y}$, MAR given $\boldsymbol{Z}$, or MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$. In the latter two cases, $\boldsymbol{Z}$ needs to be included in the imputation model.

sponse is included. Accordingly, IRT parameters can be estimated unbiasedly from the incomplete data matrix if the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y}$. This was demonstrated by Glas (2006) using data from computerized adaptive testing. Given the missing data mechanism w.r.t. $\boldsymbol{Y}$ is MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$ or only given $\boldsymbol{Z}$, the covariates need to be included in the estimation of the measurement model. For example, a routing test can be included in a latent regression model (LRM) or a multiple group IRT model (e. g. DeMars, 2002). These approaches can be seen as method-based approaches for item nonresponses. All of these methods are well studied and appropriate when the missing data mechanism is ignorable (e.g. Allison, 2001; Little & Rubin, 2002; Rubin, 1976; Schafer, 1997). For that reason, they were not discussed in detail here in this work. However, these approaches are not sufficient for nonignorable item nonresponses.

More recently, MIRT models for nonignorable item nonresponses have been introduced (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). Here, the derivation of these MIRT models from selection models (SLM) (Allison, 2001; Dubin & Rivers, 1989; Heckman, 1976, 1976; Little & Rubin, 2002; Little, 2008) was demonstrated with emphasis on the underlying assumptions. Starting from these MIRT models, further models were developed such as the LRM and the MG-IRT model for item nonresponses that are NMAR. The latter was proposed by Rose et al. (2010). In this chapter, the underlying rationale of LRM and the MG-IRT models was outlined in detail. These two model-based approaches are proposed as more parsimonious alternatives to MIRT models since the measurement model of the latent response propensity can be left out. Nevertheless, the LRM and the multiple group IRT model for missing responses were derived starting from the MIRT model, which highlight the close relation between the different models.

MIRT models for item nonresponses As shown in Section 4.5.1, a joint model of $(\boldsymbol{Y}, \boldsymbol{D})$ needs to be estimated when the missing data mechanism w.r.t. $\boldsymbol{Y}$ is NMAR. In MIRT models for nonignorable missing data, it is assumed that a latent response propensity $\boldsymbol{\theta}$ is a function $f(U)$ of the person variable U exists which determines the item response propensities $P(D_i = 1 \,|\, \boldsymbol{\theta})$. Typically, $\boldsymbol{\theta}$ is assumed to be unidimensional. This is a very strong assumption that can potentially be wrong in real applications. In Section 4.5.3.4 it was demonstrated that MIRT models can fail to correct for nonignorable missing data if the multidimensionality of $\boldsymbol{\theta}$ is ignored. Here it is argued that the dimensionality of $\boldsymbol{\theta}$ needs to be studied carefully in real application. In contrast to the items Y_i, the response indicators D_i are not deliberately generated items of a rationally constructed test. There-

fore, exploratory methods are recommended to check the number of latent dimensions required to model $\boldsymbol{D}$ appropriately. In addition to theoretical considerations, exploratory analyses such as item factor analysis might help in finding the appropriate model for $\boldsymbol{D}$.

For applied researchers it might be confusing that different between- and within-item MIRT models for nonignorable missing data have been proposed (Holman & Glas, 2005; Rose et al., 2010). Which model should be preferred in a concrete application? In the context of MIRT models for nonignorable missing data, the two classes of models differ in the construction of the latent variables. The latent ability variable $\boldsymbol{\xi}$, however, is constructed equivalently in all alternative models. It is essential that the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ needs to be preserved within the joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$). Otherwise, the resulting model is not suited to correct for missing data but is a completely new model with different item and person parameters that no longer represent the parameters of theoretical interest. However, the latent variable $\boldsymbol{\theta}$ can be constructed differently. This fact was highlighted by the different notations $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}$ that refer to different MIRT models. Only in the B-MIRT model can the latent variable $\boldsymbol{\theta} = \theta_1, \ldots, \theta_P$ be interpreted as a multidimensional latent response propensity. Different specifications of W-MIRT models for nonignorable missing data have been introduced in this dissertation. It was shown that $\boldsymbol{\theta}^* = \theta_1^*, \ldots, \theta_P^*$ can be defined as a latent difference variable, so that $\theta_l^* = \theta_l - \sum_{m=1}^{M} \xi_m$, or as a latent residual $\tilde{\boldsymbol{\theta}} = \tilde{\theta}_1, \ldots, \tilde{\theta}_P$ with $\tilde{\theta}_l$ the residual of the regression $E(\theta_l \,|\, \boldsymbol{\xi})$.

Depending on the choice of the parametric model (1PLM or 2PLM) and the dimensionality of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, different B-MIRT and W-MIRT models result. In order to classify the models developed here, a general matrix equation of the logits of all manifest variables Y_i and D_i was introduced, given by $\boldsymbol{l}(\boldsymbol{Y}, \boldsymbol{D}) = \boldsymbol{\Lambda}(\boldsymbol{\xi}, \boldsymbol{\theta}) - (\boldsymbol{\beta}, \boldsymbol{\gamma_0})$. All the different models can be differentiated due to the matrix $\boldsymbol{\Lambda}$ of the item discrimination parameters. $\boldsymbol{\Lambda}$ can be divided into four blocks. The upper left block $\boldsymbol{\alpha}$ is a $I \times M$ matrix containing the item discrimination parameters for the model equations of Y_i. The upper right block is necessarily a $I \times P$ zero matrix in all considered B- and W-MIRT models. This is essential to ensure the equivalent construction and, therefore, the equivalent meaning of $\boldsymbol{\xi}$. The lower left block $\boldsymbol{\gamma}_{\boldsymbol{\xi}}$ is an $I \times M$ matrix of item discrimination parameters that specifies the conditional regressive dependencies between $\boldsymbol{D}$ and $\boldsymbol{\xi}$ given $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$ or $\tilde{\boldsymbol{\theta}}$. In turn, the lower right block $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$ refers to the $I \times P$ matrix of item discrimination parameters that specifies the conditional regressive dependencies between $\boldsymbol{D}$ and $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$ or $\tilde{\boldsymbol{\theta}}$ given $\boldsymbol{\xi}$. Table 4.13 shows how the different models can be distinguished based on characteristics of $\boldsymbol{\Lambda}$ and associated constraints.

Which of these models should be used in application? In addition to theoretical consid-

Table 4.13: Classification of 1PL- and 2PL-BMIRT and WMIRT Models for Item Nonresponses Based on the Matrix of Discrimination Parameters ($\mathbf{\Lambda}$).

Model	$\mathbf{\Lambda}$	Constraints
General	$\begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \boldsymbol{\gamma}_{\xi} & \boldsymbol{\gamma}_{\theta} \end{pmatrix}$	
MIRT Rasch models		
B-MIRT Rasch model	$\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$	$\alpha_{im} \in \{0, 1\}$ and $\gamma_{il} \in \{0, 1\}$
W_{Dif}-MIRT Rasch model	$\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \boldsymbol{\gamma}^{*}_{\xi} & \mathbf{1} \end{pmatrix}$	$\gamma^{*}_{im} = \sum_{l=1}^{P} \gamma_{il} \Rightarrow \gamma^{*}_{im} \in \{0, 1, \ldots, P\}$
Rasch-equivalent W_{Res}-MIRT model	$\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \tilde{\boldsymbol{\gamma}}_{\xi} & \mathbf{1} \end{pmatrix}$	$\tilde{\gamma}_{im} = \sum_{l=1}^{P} \gamma_{il} b_{lm} \Rightarrow \tilde{\gamma}_{im} \in \mathbb{R}$, with $\forall(l, m)\ \ Cov(\xi_m, \theta_l) = 0$
2PL-MIRT models		
2PL-BMIRT Model	$\begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\gamma}_{\theta} \end{pmatrix}$	$\alpha_{im} \in \mathbb{R}$ and $\gamma_{il} \in \mathbb{R}$
2PL-W_{Dif}MIRT Model	$\begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \boldsymbol{\gamma}^{*}_{\xi} & \boldsymbol{\gamma}_{\theta} \end{pmatrix}$	$\gamma^{*}_{im} = \sum_{l=1}^{P} \gamma_{il} \Rightarrow \gamma^{*}_{im} \in \mathbb{R}$
2PL-W_{Res}MIRT Model	$\begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} \\ \tilde{\boldsymbol{\gamma}}_{\xi} & \boldsymbol{\gamma}_{\theta} \end{pmatrix}$	$\tilde{\gamma}_{im} = \sum_{l=1}^{P} \gamma_{il} b_{lm} \Rightarrow \tilde{\gamma}_{im} \in \mathbb{R}$, with $\forall(l, m)\ \ Cov(\xi_m, \theta_l) = 0$

erations, different goodness-of-fit (GoF) statistics could be used as a decision aid to find the most appropriate model. GoF indices rest upon on the discrepancy between observed data and the model implied joint distribution of the variables in the model. Here it was argued that GoF statistics are only one criterion to decide between alternative models. A distinctive characteristic of models for nonignorable missing data is that they consist of two parts: (a) the model of substantive interest, which reflects theoretical hypothesis (target model), and (b) a model that corrects for potential biases due to missing data. It is not sufficient to consider GoF statistics in these models. The bias reduction is also an essential criterion. Furthermore, the target model needs to be preserved in the joint model of $(\boldsymbol{Y}, \boldsymbol{D})$ or $(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$ if covariates are also part of the target model. In IRT models, that is the measurement model of $\boldsymbol{\xi}$ based on Y with the parameter vector $\boldsymbol{\iota}$. To preserve the target model means that the inclusion of a model for $\boldsymbol{D}$ affects neither the construction of $\boldsymbol{\xi}$ nor the parameter vector $\boldsymbol{\iota}$, which becomes a sub-vector in the parameter vector $(\boldsymbol{\iota}, \boldsymbol{\phi})$ of the joint model. Based on these considerations, the term model equivalence in models for missing data was adapted. Typically, two or more models are regarded to be equivalent if goodness of fit statistics indicate the same model fit (Raykov & Penev, 1999; Stelzl, 1986). However, in more precise terms, equality of fit indices is just the consequence of model equivalence. As Stelz (1986), in the context of SEM, noted, models are equivalent when they imply the same variance-covariance structure. Generally, models can be regarded to be equivalent if the model-implied joint distribution of manifest variables is equal. However, considering the principal objective of using MIRT models for missing data, the equality of model implied distributions is insufficient to regard two or models as equivalent. Additional criteria have been included. In this thesis, it was proposed to consider two or more models for item nonresponses to be equivalent if three criteria are fulfilled: (a) the latent variable $\boldsymbol{\xi}$ is constructed equivalently, (b) the bias due to item non-responses is reduced equivalently, and (3) the models imply the same distribution of manifest variables $(\boldsymbol{Y}, \boldsymbol{D})$ and, therefore, have the same model fit. If these three criteria are met, none of these models are superior with respect to the accuracy of parameter estimates of the target model, here the measurement model of $\boldsymbol{\xi}$.

The different W-MIRT models were rationally derived from the B-MIRT model, taking the issue of model equivalence into account. Accordingly, the resulting between- and within-item dimensional MIRT models for missing responses are equivalent with respect to the three aforementioned criteria. Differences between the models concern the latent variable $\boldsymbol{\theta}$, which is a latent response propensity in the B-MIRT models and a latent difference variable or a latent residual in the respective W-MIRT models. The 1PL- and

2PL-B-MIRT models are the easiest to specify and to apply. Additionally, the interpretation of the model parameters is rather simple, compared to W-MIRT models. Therefore, this model is recommended for real applications.

Latent regression models for item nonresponses Up to now, MIRT models with a large number of latent dimensions are computationally demanding (Cai, 2010). Missing data theory implies that correct inference in presence of nonignorable missing data requires to model $(\boldsymbol{Y}, \boldsymbol{D})$ jointly. In this work, the idea was developed to use functions $f(\boldsymbol{D})$ instead of the complete vectors $\boldsymbol{D}$ to simplify the model. Rose et al. (2010) proposed the joint estimation of the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ and latent regression models (LRM) $E(\boldsymbol{\xi} \mid \bar{D})$, with $\bar{D} = I^{-1} \sum_{i=1}^{I} D_i$ as the proportion of answered items. Hence, $\bar{D} = f(\boldsymbol{D})$. Other functions, such as the sum score $S_D = \sum_{i=1}^{I} D_i$, can be used as independent variables in a LRM. In this thesis, the underlying rationale and the assumptions of LRMs for nonignorable missing data were outlined in detail. The basic idea is the inclusion of the latent regression $E(\boldsymbol{\xi} \mid f(\boldsymbol{D}))$. The parameters $\boldsymbol{\iota}$ of the measurement model of $\boldsymbol{\xi}$ and the parameters of the latent regression need to be estimated simultaneously in a joint model. Although much more parsimonious, LRMs are closely related to MIRT models for nonignorable missing responses. The 2PL-BMIRT model served even as a starting point to derive this class of models. Another less formal way to understand LRMs is through their close relation to models for missing responses that are MAR given $\boldsymbol{Z}$. If a variable that determines the response propensities is observable and available, then it can be included in the model as a covariate. In this case, the missing data mechanism is MAR. Such auxiliary variables are widely used in FIML estimation (e. g. Baraldi & Enders, 2010; Graham, 2003). Under the assumption that there is no DIF in the items Y_i depending on the auxiliary variables, an LRM with $E(\boldsymbol{\xi} \mid \boldsymbol{Z})$ can be included with $\boldsymbol{Z}$ the auxiliary variables. If the latent response propensity $\boldsymbol{\theta}$ were known, then the missing data mechanism would be MAR given $\boldsymbol{\theta}$. The inclusion of a latent regression $E(\boldsymbol{\xi} \mid \boldsymbol{\theta})$ would be sufficient and $\boldsymbol{D}$ could be left out. Some functions $f(\boldsymbol{D})$ can be considered as fallible measures of $\boldsymbol{\theta}$ used in the latent regression. For example, given θ is unidimensional, the sum score S_D or the mean $\bar{D}$ are fallible measures of the latent response propensity. If $\boldsymbol{\theta}$ is a P-dimensional latent variable, then a single sum score is not sufficient. Either multiple sum scores S_{Dl} are used as proxies of θ_l or person parameter estimates $\hat{\boldsymbol{\theta}} = \hat{\theta}_1, \ldots, \hat{\theta}_P$ are taken as independent variables in a multiple regression $E(\boldsymbol{\xi} \mid \hat{\boldsymbol{\theta}})$. In application it is essential to find an appropriate function $f(\boldsymbol{D})$ that summarizes the information of $\boldsymbol{D}$ appropriately. Hence, even if the LRM is used instead of MIRT models, then the appropriate

model for $\boldsymbol{D}$ should be investigated in an initial step. If a latent response propensity is assumed, then the dimensionality of $\boldsymbol{\theta}$ should be carefully studied. Typically, it is easy to obtain estimates $\hat{\boldsymbol{\theta}}$ in this stage of the analysis that can be used in subsequent analyses. The major advantage of LRMs for nonignorable missing data is the reduced model complexity compared to MIRT models. The concurrent estimation of the measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D}$ is avoided. Another benefit is the flexibility. For example, nonlinear relations between the functions $f(\boldsymbol{D})$ and ξ_m can be modeled, including polynomials $f(\boldsymbol{D})^r$ in the regression model. Finally, here it was argued that LRMs are the methods of choice for item nonresponses resulting from not-reached items. The disadvantage is the unreliability of the functions $f(\boldsymbol{D})$ if they are proxies of latent response propensities θ_l. The question is whether and to which extent the estimations of ι and $\boldsymbol{\xi}$ are affected by the unreliability, if, for example $\hat{\boldsymbol{\theta}}$, S_D or $\bar{D}$ are used as functions $f(\boldsymbol{D})$. This question should be addressed in future research.

Multiple group IRT models for item nonresponses The use of the LRMs as an alternative for MIRT models for nonignorable missing data is limited by the restricted number of available software that allow for concurrent estimation of measurement models and latent regression models. M*plus* and ConQuest can be used for applications of LRMs. However, many traditional IRT programs as BILOG-MG (Zimowski et al., 1996), PARSCALE (Muraki & Bock, 2002), and MULTILOG (D. M. Thissen et al., 2003) do not allow for LRMs. Additionally, these programs can only estimate unidimensional IRT models. Even if such programs are used for item calibration, nonignorable missing responses can be taken into account if appropriate *discrete* functions $f(\boldsymbol{D})$ can be found that serve as grouping variables in multiple group IRT models. Rose et al. (2010) applied MG-IRT models to account for nonignorable missing item responses first. In this work, the underlying rationale for this approach was given and strong relation to MIRT models and LRM for nonigorable missing data were shown. Instead of specifying a latent regression $E(\boldsymbol{\xi} \mid f(\boldsymbol{D}))$, the regressor $f(\boldsymbol{D})$ is discrete or stratified to form groups in MG-IRT models. The item parameters are constrained to be equal across the groups/strata, while the distributions of $\boldsymbol{\xi}$ can vary across the groups. The MG-IRT model for nonignorable missing data allows for heterogeneous variances and covariance structures as well as nonlinear relations between $\boldsymbol{\xi}$ and $f(\boldsymbol{D})$. The disadvantage of this approach is once more the unreliability of the discrete functions $f(\boldsymbol{D})$ if used as a proxy of a latent response propensity. Especially if $\boldsymbol{D}$ can be appropriately modeled by a multidimensional latent response propensity $\boldsymbol{\theta}$, it might be challenging to find an appropriate discrete function $f(\boldsymbol{D})$ for the

MG-IRT model. In cases of a undimensional latent response propensity θ, the estimates $\hat{\theta}$ or S_D can be used for stratification. The number of strata depends on the sample size and practicability. However, the approach is sensitive to the choice of $f(\boldsymbol{D})$ and the number of groups respectively.

Combined models for omitted and not reached items As outlined in Section 4.5.6, the assumption of mutual local stochastic independence of the response indicators D_i given the latent variables will not hold in the case of not-reached items. Therefore, MIRT models for nonignorable missing data are appropriate for item nonresponses due to omissions but inappropriate for item nonresponses due to not-reached items. However, it is reasonable to assume that both - omitted and not-reached items - result in non-ignorable missing data. A joint model for omitted and not reached items has been developed here. The basic idea is to define two vectors of indicator variables $\boldsymbol{D^{(O)}} = D_1^{(O)}, \ldots, D_I^{(O)}$ and $\boldsymbol{D^{(N)}} = D_1^{(N)}, \ldots, D_I^{(N)}$. $D_i^{(N)} = 1$ indicates that item i was reached by the test taker, and $D_i^{(N)} = 0$ indicated that the item was not reached. $D_i^{(O)} = 1$ indicates that item i was *not* omitted, whereas $D_i^{(O)} = 0$ means that item i was omitted. An item will be responded to if it is reached and not omitted by the test taker ($D_i = D_i^{(O)} \cdot D_i^{(N)}$). Unfortunately, the vector $\boldsymbol{D^{(O)}}$ will also suffer from missing data some items are not reached. The missing data mechanism w.r.t. $\boldsymbol{D^{(O)}}$ can also be MCAR, MAR, or NMAR. It was shown that item nonresponses due to not-reached items can appropriately be modeled by latent regressions models $E[\boldsymbol{\xi} \,|\, \boldsymbol{D^{(N)}}]$. In the same way, missingness in $\boldsymbol{D^{(O)}}$ can be taken into account by an LRM with $E[\boldsymbol{\theta} \,|\, \boldsymbol{D^{(N)}}]$, where the measurement model of $\boldsymbol{\theta}$ is given by $\boldsymbol{D^{(O)}}$. The combination of an MIRT model of $(\boldsymbol{\xi}, \boldsymbol{\theta})$ based on $(\boldsymbol{Y}, \boldsymbol{D^{(O)}})$ with the multivariate latent regression $E[(\boldsymbol{\xi}, \boldsymbol{\theta}) \,|\, \boldsymbol{D^{(N)}}]$ accounts for both - omitted and not-reached items. Fortunately, all information of $\boldsymbol{D^{(N)}}$ is already given by the sum $S^{(N)} = \sum_{i=1}^{I} D_i^{(N)}$, which is simply the number of reached items given all test takers answered the same items in the same order. The reason is that the number of not-reached items might strongly depend on the order of processing test items. Test takers who start to answer easy items will potentially reach a higher number of items than those who prefer to begin with difficult items. Given that multiple test forms, such as different test booklets h, are administered, indicator variables I_h of the test form or booklet should be included as moderator variables in the regressions $E(\boldsymbol{\xi} \,|\, S^{(N)}, I_h)$, and $E(\boldsymbol{\theta} \,|\, S^{(N)}, I_h)$. Alternatively, a multiple group model can be used with the test form or booklet as a grouping variable. Unfortunately, in most paper-and-pencil tests it cannot be ruled out that test takers choose the order of items by themselves. In this case, the valid identification of not-reached items becomes difficult. However, this is not

a problem of the model used to deal with not-reached items and can only be solved by the test design, the type of data collection and documentation. For example, computerized testings allow for registering information, such as the order of answered items and the end of the test session. This allows for valid identification of not-reached items.

5 General Discussion

IRT based scaling procedures have become state-of-the-art in educational large scale assessments and are increasingly popular in many psychological tests with categorical items. Item nonresponses due to omitted or not-reached items occur in almost every application, especially in timed tests and low-stakes assessments. In the early 1970s, Lord showed that incorrect answer substitution (IAS) is inappropriate. Alternative methods, such as partially correct scoring (PCS), have been proposed. Since the late 1990s, multidimensional IRT models have been developed to account for item nonresponses that are NMAR (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). In fact, there is strong empirical evidence that missing responses are nonignorable in many applications. The tendency to omit items or not to reach the end of a test is often substantially correlated with indicators of persons' proficiency, which is intended to be measured (e. g. Culbertson, 2011, April; Rose et al., 2010). As Enders (2010) stated, models for nonignorable missing data rest upon strong and often untestable assumptions, discouraging applied researchers to use them. They rather prefer to assume that the missing data mechanism is MAR to justify the use of FIML or multiple imputation - the missing data methods currently considered as state-of-the-art (Schafer & Graham, 2002). However, it is difficult to decide which assumption is more critical: the assumption of an ignorable missing data mechanism, or the model assumptions of a model that account for nonignorable missing data. Of course, there is no ultimate answer to this question. However, especially in educational and psychological measurement, if test performance and missingness are substantially related, then it seems implausible to assume that missingness depends merely on observable item and test scores instead of the latent ability needed to answer the test. However, the latent ability of interest is always missing and the missing data mechanism with respect to the test items is then NMAR. With MIRT models for nonignorable missing data, a class of appropriate but rather complex models has been introduced to handle item nonresponses in IRT-based measurement models. Surprisingly, IRT parameter estimates were found to be pretty robust against missing responses that are NMAR (e. g. Pohl et al., 2011, September). The need to account for nonignorable

missing data is at hand from the theoretically point of view, but seemed not to be required from the practical standpoint, at least if IRT models are used. In fact, simply to ignore even nonignorable item nonresponses results in much less biased parameter estimates than IAS (e. g. Culbertson, 2011, April; Rose et al., 2010). Nevertheless, IAS and other ad-hoc methods are still commonly used even in prestigious large scale assessments, such as PISA. This thesis tried to answer different questions. First, is there a need for model-based approaches for item-nonresponses? Second, why not use ad-hoc methods, such as IAS or PCS instead of complex MIRT models for nonignorable missing data? Finally, the IRT model-based approaches for nonignorable missing data were considered in detail. The underlying assumptions of these models were explicitly considered and a common framework was given. Hence, the presented thesis consists of three major parts: (a) theory, (b) analyses of the impact of item nonresponses to item and person parameter estimates in psychological and educational measurement, and (c) the examination and further development of model-based approaches for missing nonresponses.

In the theoretical part, the missing data mechanisms were defined in the context of psychological and educational measurement following Rubin's taxonomy (1976). In the second part, the impact of missing data to different item and person parameter estimates was demonstrated in order to motivate the further development of missing data methods in the third part. Ad-hoc methods and model-based methods were considered. Following Huisman (2000), IAS or PCS are considered as naive imputation methods that were examined here in light of modern missing data theory and elaborated imputation methods. Subsequently, the nominal response model was studied with respect to its suitability to handle item nonresponses. Finally, MIRT models were scrutinized and further developed. Latent regression models and MG-IRT models were proposed as simpler alternatives to complex MIRT models. A common framework of these models was introduced, taking issues of model equivalence into account. The relationship between the alternative models has been outlined in detail. Strengths and weaknesses of the different models were discussed. Additionally, it was shown how these models can be combined in order to account for both omitted and not-reached items, even in complex item and test designs.

In this chapter, a short summary of the most important results will be given. Advantages and limitations of the different approaches will be discussed and recommendations for applied researchers will be given. Finally, remaining questions and unsolved problems are outlined that should be addressed in future research.

5.1 Summary and Conclusions

In Chapter 2 the classification of missing data introduced by Rubin (Rubin, 1976) was adapted to the context of psychological and educational measurement. Rubin distinguished between three different missing data mechanisms: (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) not missing at random (NMAR). In the latter case, the missing data are also termed nonignorable, whereas missing data that are MCAR or MAR are called ignorable. The terms informative and noninformative missing data are sometimes used alternatively. Missing data that are MCAR or MAR are called noninformative since missingness itself does not provide additional information about parameters of interest over and above the observable variables. For that reason, missingness is ignorable. In contrast, if the missing data mechanism is NMAR, then missingness provides additional information with respect to the parameters aimed to be estimated from sample data. This information needs to be included in parameter estimation to ensure unbiased parameter estimation and valid statistical inference. Hence, missing data are informative if the missing data mechanism is nonignorable.

In most educational and psychological assessments it is distinguished between items that constitute the measurement model of a latent variable and covariates, such the socioeconomic status and other background variables. Due to the distinction of the manifest variables in $\boldsymbol{Y} = Y_1, \ldots, Y_I$, the vector of test items, and $\boldsymbol{Z} = Z_1, \ldots, Z_J$ the multivariate covariate, three different MAR conditions were distinguished. Hence, five missing data mechanisms were defined twofold: (a) with respect to single items Y_i and (b) for the complete response vector $\boldsymbol{Y}$. The reason is that the item nonresponses of an item i can be MCAR while MAR or even NMAR for another item $j \neq i$. Accordingly, the three missing data mechanisms MCAR, MAR, and NMAR were defined with respect to single items Y_i in a first step. In a second step, these definitions were used to define the missing data mechanisms regarding the complete item vector $\boldsymbol{Y} = Y_1, \ldots, Y_I$.

The definitions of the missing data mechanisms are based on unconditional and conditional stochastic dependency between the following random variables: (a) the items Y_i and the response vector $\boldsymbol{Y}$ respectively, (b) the response indicators D_i that constitute the vector $\boldsymbol{D} = D_1, \ldots, D_I$, and (c) the covariate $\boldsymbol{Z}$. The latter is assumed to be completely observable. Altogether, five different missing data mechanisms have been proposed Y_i and $\boldsymbol{Y}$: (a) MCAR, (b) MAR given $\boldsymbol{Y}$, (c) MAR given $\boldsymbol{Z}$, (d) MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$, and (e) NMAR. This classification is reasonable since typical examples exist for each missing data mechanism. Furthermore, the methods to handle nonresponses differ between the missing data mech-

anisms. Table 5.1 gives and overview of the defined nonresponse mechanisms, including examples and proper missing data handling methods respectively. Note that the list of methods in the last column of Table 5.1 is by no means exhaustive. For example, special multiple imputation approaches have also been proposed for nonignorable missing data (Durrant & Skinner, 2006; Rubin, 1987). However, since MI is currently almost exclusively used as a data augmentation method for ignorable missing data, it is not listed as a method for item nonresponses that are NMAR. Instead of providing a complete overview of missing data methods, Table 5.1 serves primarily as a summary of suited approaches considered here in this work.

In IRT measurement models, latent variables are constructed based on manifest variables $Y_1, \ldots, Y_I$. In the last section of Chapter 2, the implications of the different missing data mechanisms with respect to the distribution of true score variables τ_i and latent variables ξ were studied. It could be shown that test takers who answer an item and those who do not complete this item differ systematically with respect to their true scores and latent ability if the missing data mechanism is NMAR. The results imply that potentially each item is answered by a different sub-sample that is representative for a different population with respect to the distribution of the latent variable. Especially more difficult items are more likely skipped by persons with lower ability levels. Hence, these items are completed by test takers with, on average, higher ability levels. This might be especially problematic in norm-referenced assessment based on CTT but can also cause biased parameter estimation in IRT models. The impact of item nonresponses to item and person parameter estimates was analyzed in detail in Chapter 3.

Bias of item and person parameter estimates due to item nonresponses From missing data theory follows that ML and Bayesian inference is invalid if the missing data mechanism is nonignorable, unless a model for missingness represented by $\boldsymbol{D}$ is included in parameter estimation. Surprisingly, results of real data analyses suggested that IRT parameters are pretty robust against nonignorable missing data (Pohl et al., 2011, September; Rose et al., 2010). It was repeatedly found that the use of ad-hoc methods, such as IAS or PCS, result in even more biased parameter estimates than ignoring item nonresponses that are NMAR. Therefore, the question was raised whether IRT models are robust enough simply to ignore missing data even if the nonresponse mechanism is actually nonignorable. In this case, neither ad-hoc methods nor complex model-based approaches would be required. The bias of sample estimates of item difficulties, item discriminations, and different person parameter estimates (Sum score, proportion correct

Table 5.1: Overview of Missing Data Mechanisms with Typical Examples and Potential Solutions.

Missing data mechanism	**Example**	**Appropriate missing data methods**
MCAR	Planned missing by design (i.e. balanced incomplete block designs and multimatrix sampling with randomly assigned test booklets)	Item nonresposnes can be ignored. Even listwise deletion is allowed. However, to increase efficiency, multiple imputation can be used.
MAR given $\boldsymbol{Y}$	Computerized adaptive testing (CAT) with fixed starting items or randomly chosen initial items	Item nonresposnes can be ignored in JML and MML estimation. Multiple imputation might increase efficiency in item parameter estimation from CAT data.
MAR given $\boldsymbol{Z}$	Two-stage testing using background variables or routing tests ($\boldsymbol{Z}$) to determine the assigned test form $\boldsymbol{Y}$.	Using a joint model for $(\boldsymbol{Y}, \boldsymbol{Z})$ e.g. latent regression model with $E(\boldsymbol{\xi} \mid \boldsymbol{Z})$, or multiple group IRT models (for discrete or categorized continuous $\boldsymbol{Z}$). Alternatively, multiple imputation can be used with $\boldsymbol{Z}$ in the imputation model.
MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$	CAT using background variables or routing tests ($\boldsymbol{Z}$) to determine the start items of the actual test ($\boldsymbol{Y}$).	Joint model for $(\boldsymbol{Y}, \boldsymbol{Z})$ e.g. latent regression model with $E(\boldsymbol{\xi} \mid \boldsymbol{Z})$, or multiple group IRT models (for discrete or categorized continuous $\boldsymbol{Z}$). Alternatively, multiple imputation that requires both $\boldsymbol{Z}$ and $\boldsymbol{Y}$ in the imputation model.
NMAR	The probability of item nonresponses depends on persons proficiency and, therefore, on the latent variable $\boldsymbol{\xi}$.	A joint model for $(\boldsymbol{Y}, \boldsymbol{D})$ is required. Omitted items can be controlled using MIRT models such as B-MIRT and W-MIRT Rasch models, 2PL-BMIRT -, 2PL-W$_{Dif}$MIRT -, and 2PL-W$_{Res}$MIRT model, latent regression models $E[\boldsymbol{\xi} \mid f(\boldsymbol{D})]$, or multiple group IRT models with groups formed by discrete functions $f(\boldsymbol{D})$. Not-reached items can be handled by latent regression models and MG-IRT models. Combinations of the models can be used (i.e. MIRT model with LRM for omitted and not-reached items).

score, ML-, Weighted ML-, and EAP estimates) due to missing data were studied analytically and empirically to highlight the need for appropriate missing data methods for nonignorable item nonresponses. The purpose of the detailed bias analysis was threefold. First, it should be demonstrated that different measures typically used in psychological and educational measurement, such as the sum score S and the proportion correct score P^+, are affected quite differently by missing data depending on the missing data mechanism. Second, analytical examinations of missing induced biases is easy in observed test scores S and P^+ and the expected values $E(Y_i)$, such as population specific measures of item difficulty. The findings from these analytical considerations have been used to derive hypotheses about biasedness of IRT-based person and item parameter estimates, which is difficult to investigate by analytical means. Third, the extent of the bias of IRT parameter estimates was studied using a simulation study.

In many test applications, the sum score or number correct score S is used as a test score to quantify persons' characteristics of interest. Here it was demonstrated that S will be negatively biased under any missing data mechanism. The reason is that the use of the number correct score is identical to scoring missing responses as incorrect or $Y_i = 0$ respectively. Thus, there is an implicit missing data scoring when S is used that introduces biases even if the missing data mechanism is MCAR. More formally, it could be demonstrated that the sum score under any missing data mechanism S_{Miss} is equal to the sum score of product variables $Y_i \cdot D_i$. Hence, S_{Miss} is a new random variable different from S. Apart from distributional differences, both variables differ in their meaning. S_{Miss} combines two pieces of information: (a) the test performance and (b) the ability or willingness to respond to test items. Hence, S_{Miss} is not purely a measure of test performance but reflects other persons or design characteristics as well. The variance of the sum score contains construct-irrelevant variance, jeopardizing test fairness as well as the validity of the number correct score. The sum score is essentially equivalent to incorrect answer substitution still commonly used to handle item nonresponses in IRT measurement models. The findings imply that IAS means to replace the items Y_i by the product variables $Y_i \cdot D_i$. As a consequence, the latent variable in one- and two-parameter IRT models are constructed differently, which in turn affects the interpretation of person parameters. Under IAS, the latent variable is a linear combination of the latent variable of interest and the latent response propensity. The results highlight impressively that missing data and inappropriate methods to handle them are a thread of validity.

At first glance, the proportion correct score P^+ seems to overcome the problem of the number correct score in presence of missing data, because P^+ can be seen as a stan-

dardized number correct score, where S is standardized individually by the number of answered items. However, item nonresponses due to omitted and not-reached items typically occur not randomly. A detailed analysis of item nonresponses of the PISA 2006 data revealed that more difficult items are preferably skipped, while easier items are more likely to be completed (Rose et al., 2010). Furthermore, Culbertson (2011, April) found that omission rates in items with an open response format increases with lower ability estimates. But not only the number of completed items decreases typically with lower proficiency levels, but each test taker creates his or her own test. If preferably difficult items are not answered while easier items are completed, then the whole test becomes easier. To quantify this effect, the individual mean test difficulty T_β has been introduced, which is the mean of the item difficulties of only those items that are answered by a test taker. It could be shown that P^+ is no longer comparable between test takers if T_β is correlated with the latent variable of interest. It is important to note that stochastic independence between T_β and $\boldsymbol{\xi}$ is necessary but not sufficient to ensure comparability of P^+ between test takers. The only sufficient condition is the equality of item difficulties $\beta_i = \beta_j$ for all items i and j. Hence, although the proportion correct score accounts for item nonresponses, the comparability is only ensured if all items are equal with respect to item difficulty. Otherwise, P^+ is not comparable across the different test forms, which test takers implicitly create by item nonresponses.

Although the number correct score and the proportion correct score seem to be closely related, the bias patterns found in both are quite different. Whereas the number correct score is always negatively biased under any missing data mechanism since $S_{Miss} \leq S$, P^+ can also be positively biased when preferably easy items are proceeded while difficult items are skipped. In both cases, the bias is stochastically dependent on the latent variable $\boldsymbol{\xi}$ when the tendency to omit items is also correlated with $\boldsymbol{\xi}$. This implies that missing data and the way to deal with it are very critical with regard to test fairness. Whereas nonresponses are always penalized using the number correct score, omissions of difficult items are beneficial when the proportion correct score is used. Due to these differences, it can be concluded that test takers with missing data are potentially penalized or privileged, depending on the choice of the test score. Due to these findings, neither S nor P^+ can be recommended as test scores under any missing data mechanism.

The number correct score and the proportion correct score or functions of both are commonly used as person parameter estimates in tests developed based on classical test theory (CTT). It is also common to provide item parameters in CTT, such as the expected values $E(Y_i)$ as population specific measures of the item difficulty. $E(Y_i)$ is estimated by the sam-

ple means $\bar{y}_i$. In presence of missing data, the sample mean is an estimate of $E(Y_i \mid D_i = 1)$ instead of $E(Y_i)$. Since Y_i and D_i are dichotomous, stochastic independence $Y_i \perp D_i$ is necessary and sufficient to ensure that $E(Y_i \mid D_i = 1) = E(Y_i)$. If $Y_i \not\perp D_i$ and no DIF exists in items Y_i depending on the response indicators D_i, then stochastic dependence between D_i and $\boldsymbol{\xi}$ is implied. Using a simulated data example, it could be demonstrated that each item is completed by a different sample that that refers to a different subpopulations regarding the distribution of the latent variable $\boldsymbol{\xi}$. For example, in a timed test the first item i may be completed by almost all test takers, while a difficult item j at the end of the test is more likely reached and completed by test takers with on average higher proficiency levels. It is misleading to talk about a single sample in application if each item is answered by a different subsample due to a item nonresponses that are MAR or NMAR. In that case, each item can be completed by an unrepresentative subsample even if the whole sample drawn for test application was originally representative.

Although difficult to show analytically, this fact may complicate IRT item and person parameter estimation based on observable item responses $\boldsymbol{Y_{obs}}$. For example, using the EM algorithm for MML estimation, in the E-step the expected number of cases of the sample is estimated that correctly answer item i, which is used in the M-step to obtain provisional and final item parameter estimates. The effect of systematic differences between test takers who answer single items is unclear. Analytical analyses of the bias in IRT item and person parameter estimates are quite difficult. Therefore, a simulation study was used. As noted earlier, the analytical considerations of the bias found in S, P^+, and $E(Y_i)$ served for the derivations of hypotheses about the bias of sample estimates of IRT item and person parameter estimates. The study was confined to one- and two-parametric IRT models. Three-parameter models, including pseudo guessing parameters (e. g. de Ayala, 2009; Embretson & Reise, 2000), have been left out here. The bias of estimated item difficulties, item discriminations, as well as three different person parameter estimates (ML-, WML-, EAP estimates) were investigated. The conditions that systematically varied in the simulation study were: (a) the overall proportion of missing data, (b) the correlation between the tendency to process the items and the latent ability ($Cor(\xi, \theta)$), (c) the dependency between item difficulties and the mean response rate to the items, (d) sample size, and (e) the number of items Y_i in the measurement model. The conditions were chosen to emulate data constellations typically found in real applications. That is, only positive values of the correlation between the latent ability and the latent response propensity were chosen ($0 \leq Cor(\xi, \theta) \leq 0.8$). Hence, persons with higher proficiency levels have, on average, higher probabilities to complete items. Furthermore, difficult items are more likely

to be omitted than easier items, as typically found in educational testings (e. g. Rose et al., 2010). It was expected that IRT item difficulty estimates are similarly biased as the item means under the conditions used in the simulation study. Particularly, it was expected that difficult items seem to be easier since they are completed by, on average, more proficient test takers. The results of the simulation study confirmed the systematic underestimation of β_i. The extent of the bias mainly depends on the correlation between the latent ability and the response propensity and the overall proportion of item nonresponses in the data. The higher the correlation $Cor(\xi, \theta)$ and the higher the overall proportion of missing data, the more bias was found in the estimators $\hat{\beta}_i$. Both factors interact with respect to the bias. Given ξ and θ are uncorrelated, the bias of $\hat{\beta}_i$ is close to zero even for large proportions of missing data. However, the higher the correlation $Cor(\xi, \theta)$, the stronger the bias depending on the overall proportion of missing data. The sample size is also influential, albeit to a much lesser extent. With increasing sample sizes, the bias decreases. It is important to note that the results imply that $\hat{\beta}_i$ can also be positively biased if the latent response propensity and ξ are negatively correlated. However, a preference of difficult items coupled with a negative correlation $Cor(\xi, \theta)$ seems to be implausible in most real applications. Accordingly, this condition was not included in the simulation study.

Surprisingly, the bias of discrimination parameter estimates $\hat{\alpha}_i$ was only weakly dependent on the correlation $Cor(\xi, \theta)$ and the overall proportion of missing data. The most influential factor was the sample size. With $N = 500$, the discrimination parameters were on average overestimated. Only in the case of a strong correlation $Cor(\xi, \theta) = 0.8$ was a consistent negative bias of $\hat{\alpha}_i$ found, even if sample sizes were $N = 1000$ or $N = 2000$.

The systematic bias found in item difficulty estimates $\hat{\beta}_i$ suggests that person parameter estimates could be biased as well, since β_i are locations on ξ. The bias of three different IRT person parameter estimates was studied: (a) maximum likelihood (ML) estimates, (b) Warm's weighted maximum likelihood (WML) estimates, and (c) expected a posteriori (EAP) estimates. On average, ML and WML person parameter estimates were found to be negatively biased. The mean bias of the estimated item difficulties were strongly correlated with the mean bias of ML and WML person parameter estimates (ML: $r = 0.815$, WML: $r = 0.846$). Again, the overall proportion of missing data, the correlation $Cor(\xi, \theta)$, and the interaction between these two factors mainly determined the biases of person parameter estimates. As in the case of item difficulty estimates, the bias was more negative, the higher the correlation $Cor(\xi, \theta)$ and the higher the overall proportion of missing data were. Due to the interaction effect, the impact of $Cor(\xi, \theta)$ was stronger, the higher the overall proportion of missing data was. Interestingly, a slightly positive but consistent

bias was found in ML estimates when preferably more difficult items were omitted but the missing data mechanism $\boldsymbol{Y}$ was MCAR ($Cor(\xi, \theta) = 0$). This particular bias could not be confirmed for Warm's weighted ML estimates. Apart from this exception, the bias patterns of ML- and WML estimates were very similar.

In contrast, EAP person parameter estimates were affected quite differently. The mean bias of the EAPs was found to be close to zero in almost all conditions of the simulation study. However, as Bayesian estimates, EAPs suffer from the shrinkage effect. That is, the more the estimand ξ deviates from $E(\xi)$, the larger the absolute value of the expected bias. The shrinkage effect additionally increases, the less observed information is available and, therefore, the higher the proportion of missing responses is. The shrinkage effect leads to a negative correlation between ξ and the bias of the EAPs even in absence of missing data but is increased by any loss of information such as item nonresponses. The more item nonresponses occur, the stronger the effect of the prior distribution on parameter estimation, and the more the EAPs shrink toward the expected value of the prior. This is reflected by a decreased variance of the EAP estimates. From that point of view, there is a systematic bias at the individual level if the person?s value ξ differs from $E(\xi)$. This bias is considerably increased by missing data even if the missing data mechanism is MCAR. Moreover, when EAPs are used as test scores, the omission of items can be advantageous for some persons while disadvantageous for others. Especially low proficient persons tend to produce item nonresponses. The combination of skipping difficult items while responding to easy items and the shrinkage toward the mean leads to a positive bias in persons with below-average proficiency levels. In turn, persons with values of the latent variables above the expected value $E(\xi)$ show an increasingly negative bias with increasing proportions of item nonresponses. Since the EAP was, on average, nearly unbiased, the positive and negative biases cancelled each other out. Hence, the omission of items might be beneficial for some and unfavorable for others depending on the latent variable ξ and the nonresponse behavior. This is highly questionable in terms of fairness. Once more, missing data turn out to be a matter of test fairness.

Finally, the effect of missing data on the standard errors and on the standard error function, respectively, and the marginal reliability was studied. It could be shown that under any missing data mechanism, the marginal reliability is no longer a function of item parameters and the distribution of the latent variable, but depends on the missing data pattern too. Strictly speaking, there are as many standard error functions as missing data patterns $\boldsymbol{D} = \boldsymbol{d}$ exist. Hence, each value of ξ is estimated with a different accuracy depending on the missing data pattern. Since the marginal reliabilities of ML- and WML-

estimates are calculated on the basis of the standard errors, the interpretation changes. In presence of missing data, the marginal reliability is the average reliability with respect to a particular population with its specific distribution of the latent variable *and* its specific nonresponse mechanism. Hence, the same population under study assessed with the same set of items can result in quite different marginal reliability estimates if the proportion of missing data differs. The results apply for ML-, WML-, and EAP person parameter estimates.

Ad hoc methods for item nonresponses The findings with respect to the impact of missing data on sample-based person and item parameters confirmed the need for appropriate approaches to handle item nonresponses. In a short overview, existing methods for missing data were reviewed. Analysis of complete cases (listwise deletion) or available cases (e. g. pairwise deletion) cannot be recommended in most applications. Weighting procedures are appropriate in many applications. However, in measurement models inverse probability weighting seems to be appropriate in cases of unit nonresponses but - although theoretically possible - is difficult to implement. The reason is that each item response within a response pattern is required to be weighted individually. This would be the case if each item is answered by a different population in terms of the underlying distribution of $\boldsymbol{\xi}$. Additionally, the question is how to calculate such individual item specific weights in real applications. In fact, IRT models for nonignorable missing data allow to estimate such person specific item response propensities π_{ni} under certain assumptions [1]. Hence, the estimation of the weights needed for weighting procedures require model-based methods. Furthermore, estimation procedures are required that allow for weighting individual item responses rather than weighting complete response patterns.

Data augmentation methods have become popular methods among missing data handling procedures. In this thesis, the term *data augmentation methods* subsumes all approaches that complete the observed data that suffer from missing data in a first step and to apply standard methods to filled-in data sets in a second step. Recently, multiple imputation for item nonresponses in dichotomous items used in IRT measurement models has been proved to work very well even if the proportion of missing data is large (Van Buuren, 2007, 2010). Unfortunately, most of the currently implemented algorithms for MI require that the missing data mechanism $\boldsymbol{Y}$ needs to be MCAR or MAR[2]. Hence, nonignorable item nonresponses cannot be properly handled by MI. There exist further

[1] In Section 3 it was shown how π_{ni} can be used to correct item means.

[2] If the missing data mechanism $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$ or MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$, then the covariate $\boldsymbol{Z}$ needs to be included in the imputation model.

much simpler data augmentation methods than incorrect answer substitution (IAS). To score missing responses as partially correct (PCS), as proposed by Lord (Lord, 1974), can also be seen as an imputation method. Huisman (2000) denoted such methods as naive imputation methods. In this dissertation, these two methods were also objects of research. The reason is that both IAS and PCS seem to be very plausible at first sight. The simplicity and the superficial plausibility of both methods seem to be tempting for applied researchers to use them. Although often criticized, this might be the reason why both methods are still recommended (Culbertson, 2011, April) and widely used, even in prestigious large scale assessments such as PISA (e. g. Rose et al., 2010). Once again, here it was demonstrated that IAS and PCS are highly critical for at least three reasons. First, it could be shown analytically that the implicit assumptions of IAS are unlikely to hold in almost all real applications. For example, under IAS it is assumed that the probability to solve an omitted or not-reached item is zero, which implies conditional stochastic independence $Y_i \perp \boldsymbol{\xi} \,|\, D_i = 0$. Second, theoretical inconsistencies with stochastic measurement models aimed to be applied (e.g. the 1PL- and 2PL-IRT models) result with both methods, IAS and PCS. As a result, item characteristic curves of 1PL- and 2PL- and 3PL models become incompatible with ICCs implied by IAS and PCS. For instance, given the nonresponse mechanism Y_i is MCAR, a lower and an upper asymptote different from zero and one, respectively, are implied. Third, the standard maximum likelihood estimation procedures applied to the filled-in data sets do not account for the implications of the imputation model underlying IAS and PCS. Recall that IAS and PCS were regarded as single imputation methods. For example, Lord (1974) noted that PCS is asymptotically equivalent to the imputation of random draws of a binomially distributed variable with $P(Y_i = 1 \,|\, D_i = 0) = 1/c_i$. Lord stated as well that the underlying assumption is that test takers would poorly randomly choose one of the c_i response categories for each omitted or not-reached item i. Hence, conditional stochastic independence $Y_i \perp \boldsymbol{\xi} \,|\, D_i = 0$ is implicitly assumed. Conclusively, PCS implies that two alternative mechanisms underlie the item response process: (a) solving items due to the latent ability of interest and (b) guessing. Accordingly, a mixture model would fit the resulting filled-in data set properly. Instead, for each item nonresponse, the difference $1/c_i - P(Y_{ni} = 1 \,|\, \xi_n)$ is minimized. The reason is that PCS means to replace missing responses by item-specific constants, typically $1/c_i$, and to treat these imputations like item responses that result from cognitive processing instead of guessing. It was further outlined that IAS and PCS are closely related. Alterations of PCS exist where the imputations of the item nonresponses are different from $1/c_i$. Other values can be chosen. For example, De Ayala et al. (2001)

recommended to choose $P(Y_i = 1 \mid D_i = 0) = 0.5$ even for polytomous items with more than two response categories. If $P(Y_i = 1 \mid D_i = 0)$ is set to zero, then IAS results. Hence, IAS can be regarded as a special case of PCS. Therefore, the ML estimation used under IAS is also incorrect. However, the biases with respect to the item difficulties and item discriminations are quite different. With simulated data it was shown that item difficulties and item discriminations are overestimates under IAS. In contrast, with increasing proportions of missing responses item discriminations tend toward zero when PCS is applied. Since accuracy of person parameter estimates rest upon unbiased item parameter estimates, IAS and PCS lead to biased person parameter estimates as well.

However, the term *bias* might be inappropriate for the systematic deviations between the true values of $\boldsymbol{\xi}$ and the person parameter estimates $\hat{\boldsymbol{\xi}}$ using IAS. Given the imputation model is incorrect, IAS means to replace the manifest variables Y_i in measurement model by different variables (Y_i^*) with a different distribution. Furthermore, Y_i and Y_i^* refers to different random experiments, since the use of Y_i^* additionally includes the recoding of missing responses to zero. As a result, the values $Y_i^* = 0$ can indicate two different events: (a) the item is answered incorrectly or (b) the item is missing. However, the variables Y_i^* do not allow to distinguish between the two events. Hence, two pieces of information are mixed up: (a) the performance in the test expressed by the item responses Y_i and (b) the willingness/ability to answer the item i indicated by D_i. In fact, it could be shown that Y_i^* is a function $f(Y_i, D_i)$, since $Y_i^* = Y_i \cdot D_i$. As found for the sum score S_{miss}, this confusion can result in person parameters with different meanings. In fact, when a latent response propensity θ exists that determine the probability to answer an item and $Cor(\xi, \theta =\neq 0$, then the latent variable constructed in a standard one- or two-parametric IRT model based on $Y_1^*, \ldots, Y_I^*$ is a linear combination of θ and ξ. Thus, the confusion of information in Y_i^* is reflected in the constructed latent variable. The change in the meaning of the latent variable under IAS also explains the discrepancy between the relatively large marginal reliability compared to the low squared correlation $Cor(\xi, \hat{\xi})^2$ that was consistently found in simulated data (Rose et al., 2010). The findings highlight that IRT item and person parameters are fairly sensitive to item nonresponses and their treatment. Surprisingly, simply to ignore *non*-ignorable missing responses seems to be less fatal in some situations than the use of ad-hoc methods.

In summary, IAS and PCS cannot be justified theoretically and cannot be recommended for handling item nonresponses regardless of the underlying missing data mechanism. The findings here strongly support the view that neither IAS nor PCS should be applied in educational and psychological measurement.

Nominal response model (NRM) for item nonresponses Some authors (Moustaki & Knott, 2000; Moustaki & O'Muircheartaigh, 2000) proposed to consider an item nonresponse as an additional response category and to apply the nominal response model (NRM) (Bock, 1972) for parameter estimation. The basic idea is to exploit the information about the latent variable $\boldsymbol{\xi}$ indicated by the items $Y_1, \ldots, Y_I$ by simultaneously modeling the conditional probability of an item nonresponse given the $\boldsymbol{\xi}$ in a single measurement model. In this respect, this is a model-based approach. However, item nonresponses are considered to be an additional response category. Accordingly, the manifest variables Y_i are replaced by a new variables - here denoted as R_i. In the case of dichotomous items Y_i, R_i is trichotomous respectively, with $R_i = 0$ if $Y_i = 0$, $R_i = 1$ if $Y_i = 1$, and $R_i = 2$ if Y_i is missing ($D_i = 0$). To model R_i instead of Y_i has an important property in common with data augmentation methods - the data matrix $\boldsymbol{R} = \boldsymbol{r}$ used for parameter estimation does not contain missing data anymore. In this respect, the use of the NRM for item nonresponses can be regarded as a data augmentation method. Accordingly, the standard ML estimation methods for complete data can be used for estimating parameters of NRM. In contrast to IAS and PCS, however, there are two response categories indicating that test takers failed to produce the correct answer: the wrong response ($R_i = 0$) and the missing response ($R_i = 2$). Missingness is explicitly modeled, since $P(R_i = 2 \,|\, \xi) = P(D_i = 0 \,|\, \xi)$. Hence, if the probability to answer an item depends stochastically on the latent ability, then this is taken into account by the NRM. In fact, by means of simulated data it could be shown that the item parameters of a unidimensional IRT model can be estimated unbiasedly when the individual values of ξ are known. However, that is unrealistic in application. Quite the contrary, the latent variable is constructed in a measurement model rather than simply measured. To replace the variables Y_i with R_i in the measurement model potentially affects the construction of the latent variable. As in the case of IAS and PCS, this can change the meaning of the latent variable and potentially jeopardizes the validity of the test.

Recall that the variables R_i used in NRM for nonignorable missing data are functions $f(Y_i, D_i)$ of the items and the response indicators. Note the similarity to IAS, where items Y_i are replaced by the variables Y_i^* which are also functions $f(Y_i, D_i)$. In both approaches, IAS and the NRM for item nonresponses, the measurement model is constituted by new random variables that have different distributions and, more importantly, different meanings. Both variables, Y_i^* and R_i, confound two pieces of information: missingness and test performance. The NRM for nonignorable missing data is unidimensional. As in the case of IAS and PCS, it was shown that the latent variable in the NRM is also a kind of

a mixture of the ability variable ξ and the latent response propensity θ reflecting the conflation of Y_i and D_i into a single variable R_i. It could be demonstrated that the NRM for nonignorable missing data is appropriate in one particular case, if $Cor(\theta, \xi) = 1$. In other words, when the tendency to respond to each of the items Y_i merely depends on the latent ability, then the NRM yields unbiased item and person parameter estimates. However, if $Cor(\xi, \theta) = 1$, then the response indicators D_i could simply be used as additional regular items, such as the items Y_i, in a common unidimensional Birnbaum model. The implicit assumption of $Cor(\theta, \xi) = 1$ in the NRM means that the probability of an item nonresponse of a persons is completely determined by the latent ability and that the logit of D_i is linear in ξ. These assumptions are very strong and implausible in many applications. Furthermore, they cannot be tested in the NRM. Since the NRM will fail to estimate the correct item and person parameters if $Cor(\xi, \theta) < 1$, its application is quite limited and needs to be justified.

Already in the introduction of this thesis it was mentioned that the term „biased“ is ambiguous. If IAS, PCS, and NRM are used to handle item nonresponses, then the construction of the latent variables in the IRT models is changed unless the implicit assumptions of these methods are met. Hence, if the assumptions are violated, then systematic differences in numerical values of *true* parameters and parameter estimates reflect implicit theoretical differences due to changed measurement models based on different random variables rather than biased parameter estimation. Strictly speaking, different random experiments are considered. From this point of view, inappropriate missing data methods may unpredictably change the estimands rather than produce inconsistent estimates. In other words, the *true* item and person parameters under IAS, PCS, and NRM are different from those in the measurement model of ξ based on $\boldsymbol{Y}$. Nevertheless, the term bias was consistently used for differences between parameters and their estimates throughout this work. This is reasonable from the practical point of view. In order to derive practical recommendations, it does not matter at all whether the estimand changes implicitly or the estimate is biased. Both threaten test fairness and the validity of test results.

In summary, the results of the analytical considerations and simulated data examples revealed that not only missing data but also the way to handle them are potentially a threat of accuracy, reliability, test fairness, and validity. Even the meaning of the latent variable can be affected if ad hoc methods are used thoughtlessly. Furthermore, ad hoc methods such as IAS and PCS conflict with the assumptions of most IRT models aimed to be applied. Theoretical inconsistencies and distorted parameter estimates result. The NRM for nonignorable missing data is a model-based approach resting upon very strong

implicit assumptions, so that the applicability is quite limited. However, the basic idea of the NRM is to include a model for the probability of missingness of items. This is a distinctive feature of all model-based approaches for non-ignorable missing data. The IRT models for item nonresponses considered in detail in Chapter 4.5 of this thesis rest upon this fundamental idea. Nevertheless, the NRM as well as IAS and PCS cannot be recommended. The implicit assumptions underlying these approaches are unlikely to hold in most psychological and educational assessments.

MIRT models for nonignorable item nonresponses In most applications, it is assumed that the missing data mechanism $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y}$ or $\boldsymbol{Z}$ or both. This assumption is not testable in application. However, this assumption allows to choose between different methodologically-sound and well-established approaches, which are implemented in many statistical software packages. For example, multiple imputation (MI) and full information maximum likelihood (FIML) estimation are regarded as state-of-the-art methods but require that the missing data mechanism is MAR (Schafer & Graham, 2002). Suitability of these methods for measurement models has also been studied. FIML is routinely used in SEM with missing data (Arbuckle, 1996; Enders, 2001b; Enders & Bandalos, 2001). MI for categorical items is an appropriate method to handle missing responses in IRT measurement models even if the number of item nonresponses is large (Van Buuren, 2007, 2010). Furthermore, commonly used JML and MML estimations as implemented in the most statistical software can be regarded as full information ML estimators since each observed item response is taken into account. Accordingly, Glas (2006) showed that unbiased IRT parameter estimates can be obtained by simple IRT models given the missing data mechanism $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y}$. No adoptions of the estimators or the estimation algorithms are required in this case. If the missing data mechanism $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$ or $(\boldsymbol{Y}, \boldsymbol{Z})$, then the covariate $\boldsymbol{Z}$ can be included in an LRM or, in the case of discrete covariates, MG-IRT models can be applied alternatively. In other words, if the MAR assumptions are met, then many different approaches exist to handle missing responses. However, only a few methods exist for nonignorable missing data that typically rest on strong assumptions.

Although the MAR-assumption is tempting, there is strong empirical evidence that missingness is nonignorable in many applications. For example, the rates of omitted or not-reached items at the end of a test were often found to be dependent on the test performance. The latter is a fallible indicator of the latent ability, and it is reasonable to assume that the item non-responses stochastically depends on the latent ability $\boldsymbol{\xi}$, which is per se

missing. In this case, the missing data mechanism $\boldsymbol{Y}$ is very likely NMAR. Neither the currently implemented MI procedures nor FIML ensure unbiased parameter estimation in this case. Suited methods are required that account for nonignorable item nonreponses. There is a certain similarity between model-based approaches for missing data that are MAR given $\boldsymbol{Z}$ or $(\boldsymbol{Y}, \boldsymbol{Z})$ and models for nonignorable missing data. In both cases, the target model of interest, here the measurement model of $\boldsymbol{\xi}$ constituted by $\boldsymbol{Y}$, needs to be extended to include auxiliary variables $\boldsymbol{Z}$ that account for missing data. If the missing data mechanism $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$ or $(\boldsymbol{YZ})$,then a joint model for $(\boldsymbol{Y}, \boldsymbol{Z})$ needs to be specified. Similarly, a joint model for $(\boldsymbol{Y}, \boldsymbol{D})$ is required if the missing data mechanism is NMAR. As outlined by Rubin (1976), likelihood-based and Bayesian inference are conditional on $\boldsymbol{D}$ when the nonresponse mechanism is nonignorable. This implies in turn that the missing pattern $\boldsymbol{D}$ is informative with respect to the parameters of the target model. Using this information reduces the bias due to missing data. This is the rationale underlying model-based approaches for nonignorable missing data. The essential question is how to specify a common model for $(\boldsymbol{Y}, \boldsymbol{D})$ that preserves the target model as a submodel. Unfortunately, there is no unique answer for all classes of models. In general, two broad classes of models were proposed to handle nonignorable missing data - selection models (SLM) (Dubin & Rivers, 1989; Heckman, 1976, 1979; Little, 1993, 1995, 2008; Puhani, 2000; Winship & Mare, 1992) and pattern mixture models (PMM) (Little, 1993, 1995, 2008). SLMs and PMMs were originally developed outside the context of educational and psychological measurement. Multidimensional IRT models for nonignorable missing responses have been proposed to handle nonignorable missing responses as test items (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010). These models were scrutinized and further developed here in this thesis. At the beginning it was shown that these MIRT models can be regarded as SLMs with certain assumptions. However, the same models could alternatively be derived from PMMs under the same assumptions. The essential idea underlying MIRT models for nonignorable item nonresponses is to model the stochastic dependency between $\boldsymbol{Y}$ and $\boldsymbol{D}$ by a common latent variable model. The existence of a latent response propensity $\boldsymbol{\theta}$ is assumed. The vector response indicators $\boldsymbol{D} = D_1, \ldots, D_I$ constitute the measurement model of $\boldsymbol{\theta}$. The measurement model of the latent ability $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ is the target model preserved in the joint measurement model of $(\boldsymbol{\xi}, \boldsymbol{\theta})$ based on $(\boldsymbol{Y}, \boldsymbol{D})$. An important assumption in this model is the local stochastic independence of all Y_i and D_i. More specifically, it is assumed that $Y_i \perp (\boldsymbol{Y}_{-i}, \boldsymbol{D}) \,|\, \boldsymbol{\xi}$ and $D_i \perp (\boldsymbol{Y}, \boldsymbol{D}_{-i}) \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$. In application, violations of the assumptions of local independence

results often from ignored multidimensionality of the latent variable.

For example, let there be groups of items that differ with respect to the item format (multiple choice versus open-constructed) or the content (mathematics versus biology). If test takers differ with respect to their tendency to omit items between these domains, then the latent response propensity will be multidimensional. In such a case, the two-dimensional latent response propensity would underlie $\boldsymbol{D}$. However, the issue of the dimensionality of the latent response propensity is typically ignored in the literature. In most applications, the latent response propensity is assumed to be unidimensional. In this thesis, it could be demonstrated that MIRT models for nonignorable missing data potentially fail to account sufficiently for item nonresponses when unidimensionality of θ is assumed erroneously. In contrast to the items Y_i, the response indicators D_i are not theoretically constructed items. Typically, there are only ad-hoc hypotheses about the dimensionality of the latent response propensity. Therefore, here it is argued that exploratory methods, such as item factor analysis (e. g. Tate, 2003; Wirth & Edwards, 2007) and item clustering (e. g. Reckase, 2009), should complement theoretical considerations to find the appropriate model for $\boldsymbol{D}$. The assumption of unidimensionality of the latent response propensity should be studied in application.

A closer examination of existing literature of MIRT models for item nonresponses revealed that different alternative MIRT models have been proposed. In general, between-item multidimensional IRT (B-MIRT) models and within-item multidimensional IRT (W-MIRT) models for item nonresponses can be distinguished. Typically, the models are considered to be equivalent. Holman and Glas (2005) showed analytically how these models are related. However, the substantive meaning of some model parameters differs between the models. Rose et al. (2010) demonstrated the differences between one-parametric B- and W-MIRT models for nonignorable missing data. They showed that differences in the model of $\boldsymbol{D}$ result in differently constructed latent variables. A latent response propensity is only constructed in B-MIRT models. In W-MIRT models a latent difference variable $(\theta - \xi)$ is constructed. In fact, in one-parameter models with θ and ξ unidimensional each, the B- and W-MIRT model are also equivalent in terms of models fit. However, as shown here, this is not necessarily the case for two-parameter models. Furthermore, W-MIRT model for nonignorable missing data requires additional restriction for model identification. Depending on the particular restriction chosen to identify the model, different W-MIRT models result. The models differ in complexity and interpretation of model parameters. Which of these models should be used in application? Is there any benefit from using W-MIRT models instead of B-MIRT models or vice versa?

Here, it was shown how the W-MIRT models can be analytically derived starting from the B-MIRT model. This was done separately for one- and two-parameter models. The class of models was extended to include cases with a complex multidimensional structure of latent abilities $\boldsymbol{\xi} = \xi_1, \ldots, \xi_M$ and latent response propensities $\boldsymbol{\theta} = \theta_1, \ldots, \theta_P$.

Two different W-MIRT models were derived theoretically, starting with the definition of latent variables. It was shown that only in the B-MIRT model is a latent response propensity constructed. Two alternative W-MIRT models have been derived where the latent response propensity is replaced by functions $f(\boldsymbol{\xi}, \boldsymbol{\theta})$. The resulting models are structurally similar to latent change models (e. g. Steyer, Eid, & Schwenkmezger, 1997; Steyer, Krambeer, & Hannover, 2004) and recently proposed correlated-trait-correlated-method factor models (e. g. Eid et al., 2008; Geiser, Eid, Nussbeck, Courvoisier, & Cole, 2010; Geiser & Lockhart, 2012, February 6; Pohl & Steyer, 2010). In the derivation of alternative MIRT models based on $(\boldsymbol{Y}, \boldsymbol{D})$ needs to preserve the target model - the measurement model of $\boldsymbol{\xi}$. In other words, the construction of $\boldsymbol{\xi}$ needs to be unaffected, so that the person parameters and the item parameters with respect to Y_i are equal. Under this condition four different W-MIRT models could be derived. In the W_{Dif}MIRT Rasch model and the 2PL-W_{Dif}MIRT model, the latent response propensity $\boldsymbol{\theta}$ is replaced by a multidimensional latent difference variable $\boldsymbol{\theta^*} = \theta_1^*, \ldots, \theta_P^*$ with $\theta_l^* = \theta_l - \sum_{m=1}^{M} \xi_m$ (for all $l \in 1, \ldots, P$). In the W_{Res}MIRT Rasch model and the 2PL-W_{Res}MIRT model, $\boldsymbol{\theta}$ is replaced by a latent residual $\tilde{\boldsymbol{\theta}} = \tilde{\theta}_1, \ldots, \tilde{\theta}_P$ with $\tilde{\theta}_l = \theta_l - E(\theta_l \,|\, \boldsymbol{\xi})$. All these alternative models were rigorously mathematically developed starting from the B-MIRT Rasch model and the 2PL-BMIRT model respectively. This allowed for the derivation of model implied constraints with respect to the item discrimination parameters in the different one- and two parametric W-MIRT models.

A general model equation has been introduced that allows to distinguish the different MIRT models for nonignorable missing data formally (see Equations 4.79 and 4.80). The structure of the matrix $\boldsymbol{\Lambda}$ of item discrimination parameters and the constraints imposed for the single elements in $\boldsymbol{\Lambda}$ are distinctive for the MIRT models considered here (see Table 4.13). Under these constraints, B-MIRT and W-MIRT models turned out to be equivalent in terms of model fit. Hence, GoF cannot serve as a decision aid to determine the most appropriate model. The fit of a model to the data is only one criterion and, possibly, not the most important one to choose the best missing data model in a real application. For example, it is easy to specify a model for $(\boldsymbol{Y}, \boldsymbol{D})$ that is equivalent or even better in terms of model fit but practically of no use since the target measurement model is not preserved. Recall that the nonresponse model (model of $\boldsymbol{D}$) is actually a

nuisance (Enders, 2010). The only reason to include $\boldsymbol{D}$ is for the reduction or elimination of bias with respect to the parameter estimates of the target model. In IRT models, the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ with the parameter vector $\boldsymbol{\iota}$ is of crucial interest. Here it was outlined that two alternative missing data models can be regarded to be equivalent - in the sense that they are equally suited to be applied - if they equally reduce the bias in parameter estimates $\hat{\boldsymbol{\iota}}$. Common concepts of model equivalence that focus on model fit (e. g. Raykov & Penev, 1999; Stelzl, 1986) are not sufficient when missing models are considered. In this work, it was argued to consider two or more missing data models to be equivalent if three criteria are fulfilled: (a) the latent variable $\boldsymbol{\xi}$ is constructed equivalently as in the complete data model, (b) the bias due to item non-responses is reduced to the same extend, and (3) the models imply the same distribution of manifest variables ($\boldsymbol{Y}$, $\boldsymbol{D}$), and, therefore have the same model fit. Only if these three criteria are met, then none of these models are superior with respect to the quality of parameter estimates of the measurement model of $\boldsymbol{\xi}$.

The W-MIRT models rationally derived in this work have been shown to be equivalent to the respective B-MIRT models with respect to the three criteria. The parameter vector $\boldsymbol{\iota}$ is the same in all models. The vector $\boldsymbol{\phi}$ of parameters referring to the probability model of $\boldsymbol{D}$ are different, implying interpretational differences in this part of the model. Simulated data example confirmed that the 2PL-BMIRT , the 2PL-W$_{Dif}$MIRT , and the 2PL-W$_{Res}$MIRT models are equivalent in the sense defined here. The same applies for the B-MIRT-, the W$_{Dif}$-, and the W$_{Res}$MIRT Rasch model. Differences between the models exist with regard to the practicability. W$_{Res}$MIRT Rasch models and 2PL-W$_{Dif}$MIRT and 2PL-W$_{Res}$MIRT models require the specification of nonlinear constraints. Many IRT programs allow for equality constraints but only a few allow to specify complex nonlinear constraints. However, M*plus* (Muthén & Muthén, 1998 - 2010) is very flexible and allows to estimate all models considered in this work. Example input files are given in the Appendix (see 5.3). Unfortunately, the number of constraints increases rapidly with the number of items Y_i and latent dimensions ξ_m and θ_l in the model. This makes the use of 2PL-W$_{Dif}$MIRT and 2PL-W$_{Res}$MIRT models difficult. If the constraints with respect to the item discrimination parameters of the 2PL-W$_{Res}$MIRT model are simply ignored, then the relaxed 2PL-W$_{Res}$MIRT model results. This model has been proposed as an alternative model to the B-MIRT model (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999). Here it was shown that the relaxed 2PL-W$_{Res}$MIRT model is not equivalent to the 2PL-BMIRT model in terms of model fit since more model parameters need to be estimated. However, if the assumptions of the 2PL-BMIRT model are met,

then the relaxed 2PL-W_{Res}MIRT model is equivalent in terms of the construction of $\boldsymbol{\xi}$ and the bias reduction in item and person parameter estimates. In other words, the relaxed 2PL-W_{Res}MIRT model is overparameterized but equally suited to account for nonignorable item nonresponses. Despite the lack of parsimony, the advantage of this version of the model is its applicability in programs that allow for bifactor analysis, such as TESTFACT (Bock et al., 2003).

The between-item multidimensional models such as the 2PL-BMIRT Model and the B-MIRT Rasch Model are much easier to handle and do not require the specification of non-linear constraints. The interpretation of latent variables and their correlations is much easier. The latent variable $\boldsymbol{\theta}$ is a multidimensional latent response propensity instead of a function $f(\boldsymbol{\xi}, \boldsymbol{\theta})$, such as $\boldsymbol{\theta}^*$ or $\tilde{\boldsymbol{\theta}}$. Accordingly, the correlations $Cor(\xi_m, \theta_l)$ are informative with respect to the strength of the dependencies between the missingness of item responses and the person's ability . The stochastic dependencies between the items Y_i and the response indicators D_i are implied by the latent covariance structure between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. Insofar, the applications of MIRT models for nonignorable missing data are of diagnostic value. The extent to which nonresponses and ability are related under a certain test design can be studied. Of course, the same information can be extracted from W-MIRT models with difficulty. However, due to their practicability and easier interpretation, B-MIRT models are recommended as the MIRT model of choice to handle omitted responses that are NMAR.

The disadvantages of MIRT models become clear when the assumptions are considered. That is the assumption of local stochastic independence $Y_i \perp (\boldsymbol{Y_{-i}}, \boldsymbol{D}) \,|\, (\boldsymbol{\xi})$ of the items and the response indicators $D_i \perp (\boldsymbol{Y}, \boldsymbol{D_{-i}}) \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$. Due to the latter, MIRT models are not appropriate to handle not-reached items. This was shown analytically in Section 4.5.6. Furthermore, all stochastic dependencies between the items Y_i and D_i are implied by the stochastic dependencies between the latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. Hence, an appropriate model for $\boldsymbol{D}$ is a prerequisite. It was demonstrated that ignoring multidimensionality of $\boldsymbol{\theta}$ can make MIRT models for missing responses ineffective. For that reason, it was argued that the dimensionality underlying $\boldsymbol{D}$ should be carefully studied, including exploratory methods, such as item factor analysis.

In most current implementations, only linear stochastic dependencies between latent variables in MIRT models can be taken into account. A latent variance-covariance matrix of the latent dimensions or latent residuals is used to describe the unconditional and conditional multivariate normal distribution respectively. Furthermore, only linear regressions between the latent dimensions scan be specified. However, if non-linear relations

exist between dimensions ξ_m and θ_l, then the MIRT models for nonignorable missing data potentially fail to adjust for missingness.

Latent regression models and multiple group IRT model for item nonresponses An important drawback of all MIRT models examined here is their complexity. The number of manifest variables doubles when the response indicator vector $\boldsymbol{D}$ is included. Especially in large scale assessments with multimatrix sampling designs, the measurement models contain far more than a hundred items Y_i. Hence, a joint model of $(\boldsymbol{Y}, \boldsymbol{D})$ can easily comprise several hundred manifest variables. Accordingly, the number of latent variables in the model may increase as well. Given that both latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, are multidimensional, the models become computationally demanding. As Cai (Cai, 2010) noted, high dimensional IRT models remain numerically challenging. Therefore, simpler models would be helpful. Missing data theory implies that correct inference in presence of nonignorable missing data requires to model $(\boldsymbol{Y}, \boldsymbol{D})$ jointly. In this work, the idea was developed to use functions $f(\boldsymbol{D})$ instead of the complete vector $\boldsymbol{D}$. Rose et al. (2010) were the first to propose the inclusion of latent regression model (LRM) including $E(\boldsymbol{\xi} \mid \bar{D})$, with $\bar{D} = I^{-1} \sum_{i=1}^{I} D_i$. The parameters of this regression need to be estimated jointly with the parameter of the measurement model ($\boldsymbol{\iota}$). Here the underlying rationale of this approach was outlined. Each regression $E[\boldsymbol{\xi} \mid f(\boldsymbol{D})]$ can be used if an appropriate function $f(\boldsymbol{D})$ can be found. In some cases, the number of responded items $S_D = \sum_{i=1}^{I} D_i$ or the proportion of responded items $\bar{D}$ can be sufficient. If $\boldsymbol{D}$ underlies a multidimensional latent response propensity $\boldsymbol{\theta}$ with a complex structure, then individual estimates $\hat{\boldsymbol{\theta}}$ can be generated in a first step based on a model of $\boldsymbol{D}$ alone. The estimates can be used as independent variables in a LRM in the second step that includes estimation of $\boldsymbol{\iota}$. The functions $f(\boldsymbol{D})$ should be chosen as parsimoniously as possible and with the minimal loss of information. Here it was shown that in the case of 30 items, the sum score S_D used as a function $f(\boldsymbol{D})$ in a LRM results in nearly identical item and person parameter estimates (EAPs) as in the 2PL-BMIRT model. However, the number of parameters and the computational demand is considerably lower when the LRM is used.

However, theoretically the LRM for item nonresponses and the 2PL-BMIRT model are closely related. In the latter, the latent response propensity is included by the measurement model based on $\boldsymbol{D}$. If the local independence assumptions hold true and $\boldsymbol{\theta}$ is an observable variable, then the missing data mechanism $\boldsymbol{Y}$ would be MAR given $\boldsymbol{\theta}$. ML and Bayesian inference based on a joint model of $(\boldsymbol{Y}, \boldsymbol{\theta})$ would be valid and $\boldsymbol{D}$ could be ignored. Generally, covariates can be taken into account in an IRT measurement model as

independent variables in an LRM. The joint estimation of parameters of the measurement model of $\boldsymbol{\xi}$ and the latent regression $E(\boldsymbol{\xi} \mid \boldsymbol{\theta})$ using MML estimation would be equivalent to FIML estimation with auxiliary variables (Graham, 2003; Mislevy, 1987, 1988). Of course, in real applications the latent response propensity is unobservable. Therefore, here it was proposed to use estimates $\hat{\boldsymbol{\theta}}$ or other functions $f(\boldsymbol{D})$, such as S_D or $\bar{D}$, which can be considered as proxies of a latent response propensity. However, in the case of a multidimensional latent response propensity, the use of a single sum score S_D or proportion of answered items $\bar{D}$ is questionable. For that reason, the dimensionality of the latent response propensity should also be taken into account in the choice of the potentially multidimensional function $f(\boldsymbol{D})$. For example, sum scores S_{D_l} can be used, that are calculated by summing only that items D_i that indicate θ_l. Hence a multiple latent regression can be specified with several sum scores S_{Dl} as independent variables. Alternatively, the person parameter estimates $\hat{\boldsymbol{\theta}} = \hat{\theta}_1, \ldots, \hat{\theta}_P$ can be used. Since the initial analysis of the dimensionality underlying $\boldsymbol{D}$ is recommended in each case, the estimates $\hat{\boldsymbol{\theta}}$ can easily be obtained as a by-product and can further be used in a LRM. A special case is the use of the identity function $f(\boldsymbol{D}) = \boldsymbol{D}$ so that each single response indicator is included in the latent regression $E(\boldsymbol{xi} \mid \boldsymbol{D})$. If no other appropriate function $f(\boldsymbol{D})$ can be found, then this is the least restrictive LRM. However, the number of estimands in the model increases with the number of items in the measurement model, especially if interaction effects between D_i and D_j $(i \neq j)$ exist with respect to $\boldsymbol{\xi}$.

It was shown that the LRM is the method of choice to account for item nonresponses due to not-reached items. The assumption of local stochastic independence $D_i \perp (\boldsymbol{Y}, \boldsymbol{D}_{-i}) \mid (\boldsymbol{\xi}, \boldsymbol{\theta})$ in MIRT models for item nonresponses is always violated in the case of not-reached items. If all missing responses result exclusively from not-reached items, then all information about $\boldsymbol{D}$ is given by the number of reached or not-reached items since $\boldsymbol{D}$ always follows a perfect Guttman-pattern. In this case, S_D is always an appropriate function $f(\boldsymbol{D})$ for the LRM. If item nonresponses result from both, omitted and not-reached items, then more complex models, as proposed in Section 4.5.6, are required. These models are summarized below.

The major advantage of using the LRM for nonignorable missing data is the reduction of model complexity compared to the MIRT models, given suited functions $f(\boldsymbol{D})$ can be found. The concurrent estimation of the measurement model of $\boldsymbol{\theta}$ based on $\boldsymbol{D}$ is avoided. Furthermore, nonlinear relations between the $f(\boldsymbol{D})$ and ξ_m can be modeled by inclusion of polynomials and interaction terms. Given the estimates $\hat{\boldsymbol{\theta}}$ are used, non-linear relationships between the latent dimensions ξ_m and θ_l can be approximated. Further-

more, interactions between $f(\boldsymbol{D})$ and other covariates can be included. Exemplarily, it was demonstrated how to include $f(\boldsymbol{D})$ in a booklet design when the booklet (indicator variables of the booklets) moderates the dependency between missingness and the latent ability.

In the derivation of the LRM for missing responses the underlying assumptions where explicated. It was shown that $\boldsymbol{D}$ can be ignored if conditional stochastic independence $\boldsymbol{D} \,|\, \boldsymbol{Y}_{mis} \,|\, (f(\boldsymbol{D}), \boldsymbol{Y}_{obs})$ holds true. This assumption is only warranted to hold true if $f(\boldsymbol{D}) = \boldsymbol{D}$. If other functions than the identity function are used, then it is important that all information in $\boldsymbol{D}$ with respect to $\boldsymbol{Y}_{mis}$ is preserved in $f(\boldsymbol{D})$. Unfortunately this is untestable and will only approximately be achieved in real applications. However, theoretical considerations underline the importance of the deliberate choice of the function $f(\boldsymbol{D})$. Therefore, a careful examination of $\boldsymbol{D}$ should always precede the application of the LRM for item nonresponses. In some applications it may be difficult to find appropriate functions $\boldsymbol{D}$. In such cases, the applicability of the LRM is limited.

If the functions $f(\boldsymbol{D})$ can be regarded as proxies of a latent response propensity, then the impact of measurement error with respect to bias reduction remains unclear. It is well known that unreliability leads to biased regression coefficients and correlations. Little is known about the impact of unreliability in auxiliary variables with respect to bias reduction. Especially when the number of manifest variables is low, it is expected that unreliability of $f(\boldsymbol{D})$ derogates the bias reduction. Further research is needed to study the robustness and suitability of the LRM with different functions (fD) in different testing designs.

Unfortunately, the number of available software that allow for concurrent estimation of a measurement models and a LRM is limited. For example, M*plus* (Muthén & Muthén, 1998 - 2010) and ConQuest (Wu et al., 1998) can be utilized to apply LRMs for nonignorable missing data. However, many traditional IRT programs, such as BILOG (Zimowski et al., 1996), PARSCALE (Muraki & Bock, 2002), and MULTILOG (D. M. Thissen et al., 2003), do not allow for the inclusion of LRMs. Furthermore, these programs can only estimate unidimensional IRT models. Hence, neither LRM nor MIRT models for nonignorable missing data can be applied. However, multiple group IRT models can be fitted in these software packages. Rose et al. (2010) applied MG-IRT models to account for nonignorable item nonresponses. This approach is straightforward and closely connected to the LRMs for missing responses. Stratification is widely used in linear regression analysis (e. g. Quesenberry & Jewell, 1986). The MG-IRT model results if a discrete function $f(\boldsymbol{D})$ can be found, for example, by stratification of the proportion of completed items.

Indicator variables of the resulting strata can be used in ordinary linear regression models. Instead of using a latent regression $E(\boldsymbol{\xi} \mid f(\boldsymbol{D}))$, a multiple group model can be used with $f(\boldsymbol{D})$ as the grouping variable. Rose et al. (2010) stratified the mean response rate $\bar{D}$ in order to account for missing responses in the PISA 2006 data. They formed three groups so that the number of cases in each stratum were similar. The item parameters in the MG-IRT model were constrained to be equal across the strata, to ensure a common metric. The distributions of $\boldsymbol{\xi}$, however, could vary across the groups. The MG-IRT model for nonignorable missing data allows for heterogeneous variances and captures nonlinear relations between $\boldsymbol{\xi}$ and $f(\boldsymbol{D})$. Distributional differences with respect to the latent ability across the groups indicate that the missingness stochastically depends on $\boldsymbol{\xi}$. The advantage of MG-IRT models for missing responses is their simplicity and applicability even in software that allow neither for estimating MIRT models nor the inclusion of LRMs.

Theoretically, this approach is very close to pattern mixture models, where each missing pattern forms a group. Regarding the MG-IRT models as a special case of the LRMs implies that the unreliability of the functions $f(\boldsymbol{D})$ is also a potential threat in MG-IRT models. If a latent response propensity exists, then the use of a discrete function $f(\boldsymbol{D})$ with too few levels can be an oversimplification. Hence, to form the groups appropriately can be a nontrivial task. Again, the analysis of $\boldsymbol{D}$ should precede the application of the MG-IRT model for item nonresponses. If a MIRT model can be fitted to the data $\boldsymbol{D} = \boldsymbol{d}$, then the estimates $\hat{\boldsymbol{\theta}}$ can be stratified to form groups of the MG-IRT model. This is recommended especially in cases with a complex dimensional structure of $\boldsymbol{\theta}$. In general, applied researchers should be aware that the MG-IRT model is sensitive to the choice of grouping. As in the case of LRM for nonignorable missing data, further research is needed to study the robustness of the approach under different test designs.

A joint model for omitted and not-reached items Considering the local stochastic independence assumptions of MIRT models for nonignorable missing data as well as the properties of response indicators D_i revealed that MIRT models are appropriate for omitted responses but inappropriate to handle nonignorable missing responses due to not-reached items. The reason is that response indicators D_i and D_j $(i \neq j)$ indicating reached or not-reached items are deterministically dependent. The probability to answer an item $i + 1$ after the first not-reached item i is always zero and the probability to reach an item $i - 1$ prior to the first not-reached item i is always equal to one. This violates the assumption $D_i \perp (\boldsymbol{Y}, \boldsymbol{D}_{-i}) \mid (\boldsymbol{\xi}, \boldsymbol{\theta})$ of conditional stochastic independence of all MIRT models considered in this work. It was shown that LRMs are the method of choice to

handle nonigorable missing responses due to not-reached items. MIRT models however are suited for omitted responses. In most real applications, missing responses in a single item i result from both failing to reach the end of the test and omissions of items. How does one model nonignorable missing responses if omitted and not-reasched items needs to be treated differently? In Section 4.5.6 a joint model for omitted and not-reached has been developed that combines a MIRT model with an LRM. In order to distinguish between omitted and not-reached items, $\boldsymbol{D}$ was replaced by two vectors of indicator variables: $\boldsymbol{D^{(O)}} = D_1^{(O)}, \ldots, D_I^{(O)}$ and $\boldsymbol{D^{(N)}} = D_1^{(N)}, \ldots, D_I^{(N)}$. $D_i^{(N)} = 1$ indicates that item i was reached by the test taker and $D_i^{(N)} = 0$ otherwise. $D_i^{(O)} = 1$ indicates that item i was *not* omitted by the test taker and $D_i^{(O)} = 0$ otherwise. An item responses is observed if the item i is reached ($D_i^{(N)} = 1$) and not omitted by the test taker ($D_i^{(O)} = 1$). Hence $D_i = f(D_i^{(N)}, D_i^{(O)})$ and $\boldsymbol{D} = f(\boldsymbol{D^{(N)}}, \boldsymbol{D^{(O)}})$ respectively. The final model consists of a joint measurement model of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ based on $(\boldsymbol{Y}, \boldsymbol{D^{(O)}})$ and an LRM with two, potentially multivariate regressions $E(\boldsymbol{\xi} \,|\, S^{(N)})$ and $E(\boldsymbol{\theta} \,|\, S^{(N)})$. The latter is important since the vector $\boldsymbol{D^{(O)}}$ will also suffer from missing data if items in the end of the test are not reached. The model of $\boldsymbol{D^{(O)}}$ is the measurement model of $\boldsymbol{\theta}$ with the I regressions $P(D_i^{(O)} = 1 \,|\, \boldsymbol{\theta})$. $\boldsymbol{\theta}$ is the general tendency *not* to omit the items i. Items can only be completed or omitted by the test takers when they are reached in time. Not-reached items lead to missing data in both the items Y_i and the indicators $D_i^{(O)}$. Given that the number of not-reached items and the omission of items is stochastically dependent, the missing data mechanism $\boldsymbol{D^{(O)}}$ is also NMAR. The latent regression $E(\boldsymbol{\theta} \,|\, S^{(N)})$ accounts for these nonignorable missing data. Item nonresponses in $\boldsymbol{Y}$ will be appropriately taken into account by both the regression $E(\boldsymbol{\xi} \,|\, S^{(N)})$ and the joint model of $(\boldsymbol{Y}, \boldsymbol{D^{(O)}})$.

In application there are some difficulties in modeling not-reached items, since their identification can be difficult. Typically, the connected sequence of missing responses at the end of the test is assumed to be a result of failing to reach the end of the test. However, it cannot be ruled out that these items have been intentionally omitted. Furthermore, the current identification rules for not-reached items assume that all test takers answer the items in the presented order. However, in paper-and-pencil tests, test takers potentially choose the order of items by themselves. In this case, not-reached items and omitted become indistinguishable. Fortunately, the use of computerized testings allow for the registration of the order of answered items and a valid detection of both omitted and not-reached items. This potentially facilitates the model-based approaches for item nonresponses in psychological and educational testings.

5.2 Recommendations for Real Applications

Based on the results of this work and in line with previous research, different recommendations for applied researchers in the field of educational and psychological measurement can be derived. At first it is strongly recommended never to use ad hoc methods such as IAS or PCS to handle item nonresponses. Simply to ignore missing data in IRT models seems to be less harmful than using such ad hoc methods (e. g. Culbertson, 2011, April; Lord, 1974; Rose et al., 2010).

In order to find the most appropriate missing data method in a particular application, some questions should addressed. First, the appropriate approach to handle item nonresponses depends on the missing data mechanism. Therefore, the first question is, what is known about the missing responses? It needs to be kept in mind that nonresponses in a single item can result from not-reached items or omitted items, or they can be due to the design. The latter are planned missing data. If the design implies that planned missing data are ignorable, then only missingness due to omitted and not-reached items is of concern. This is typically the case in multimatrix-designs if the booklets are randomly assigned. If unplanned missing data exist, then it needs to be answered whether observable variables determine the missing pattern. This is difficult to answer in most applications. However, in CAT the missing data mechanism is MAR given $\boldsymbol{Y}$ or MAR given $(\boldsymbol{Y}, \boldsymbol{Z})$ if covariates $\boldsymbol{Z}$ are used to determine the starting items. In these cases, item and person parameters can be estimated unbiasedly based on MML estimation including $\boldsymbol{Z}$ respectively (Glas, 2006). Item imputation methods can be alternatively applied in these cases (Van Buuren, 2007, 2010). The covariates $\boldsymbol{Z}$ need to be included in the imputation model given the missing data mechanism $\boldsymbol{Y}$ is MAR given $\boldsymbol{Z}$ or $(\boldsymbol{Y}, \boldsymbol{Z})$.

If the test design does not allow to infer about the missing data mechanism, then the plausibility of the MCAR and the MAR assumptions should be questioned. Whereas the MCAR assumption can be tested (Chen & Little, 1999; Little, 1988b), no satisfying approaches exist to test the MAR assumptions. Hence, if the assumption of missing data being MCAR is not tenable, it should be deliberately decided whether the MAR assumptions are reasonable or not. Since no test is available, missingness and its relation with observed data should be scrutinized. The resulting statistics together with theoretical considerations are the basis to decide which procedure is justifiable to handle missing responses.

In order to study the plausibility of the MAR assumption, the missing pattern can be examined. It should be carefully studied which items preferably have been omitted or

not-reached. Do omissions occur more in items with certain response formats? Are there more nonresponses in items that address certain issues or topics? Are the omission rates in the items dependent on item characteristics, such as item difficulty or the position or the context in which the item was presented? Depending on the study there might be further questions that should be answered.

Furthermore, the relationship of the response rates of the test takers with other person variables can be studied. If covariates exist that are stochastically related with missingness, then the strength of these associations is important. Such covariates could be included in the parameter estimation. It should be asked then, how plausible the assumption of MAR given the covariate is. To gain a good first impression, descriptive statistics should be used that quantify the relationship between $\boldsymbol{D}$, $\boldsymbol{Y}$ and other covariates $\boldsymbol{Z}$. For example, the relationship between the proportion correct score and the proportion of omitted and not-reached items can be analyzed [3]. For example, Rose et al. (2010) found a correlation of $r = 0.33$ in the PISA 2006 data, indicating a relationship between proficiency and missingness. If covariates $\boldsymbol{Z}$ exist, then they should also be studied in their relation to the response indicators D_i. Depending on the scales of the variables Z_J in $\boldsymbol{Z}$, contingency tables, χ^2-tests, t-tests, logistic regressions, etc., and graphical procedures can be used.

It is also important to consider the existence of latent variables, which are inherently missing. If the number of item nonresponses is related to the performance in items that have been answered, then the missing data mechanism is only MAR if missingness is conditionally stochastically independent of the latent ability given the observed item responses and other covariates. Here it is argued that this seems unlikely in many applications. A relation between missingness and test performance is more likely implied by the stochastic dependency between missingness and the latent ability intended to be measured. In this case, the missing data mechanism is most likely nonignorable. If there is doubt that the missing data mechanism $\boldsymbol{Y}$ is MAR, then models for nonignorable missing responses should be applied. These methods can also be used in sensitivity analyses comparing models for missing data that are MAR and NMAR.

If missing data mechanism is assumed to be nonignorable, then an appropriate method or model needs to be chosen to handle item nonresponses. The applicability of the different model-based approaches depends on several factors, such as

1. The distinction between not-administered items (planned missing data), omitted

[3] Note, however, that the proportion correct score is itself affected by missing data and the relationship with the proportion of missing data might be biased and should only be used as a starting point for further analyses.

responses, and not-reached items.

2. The proportion of unplanned missing response per item (proportion of nonresponses in item i due to omission or not-reached items).
3. The number of items Y_i.
4. The number of items with a significant number of unplanned missing data.
5. The model complexity of the target model - the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$.
6. The complexity of the model for $\boldsymbol{D}$ and/or the availability of appropriate functions $f(\boldsymbol{D})$
7. Sample size
8. Software capabilities

The distinction between nonresponses due to not-administered, omitted, or not-reached items is essential. Here it was shown that omitted and not-reached items need to be treated differently even if both result in nonignorable missing data. It has been proposed to distinguish between $D_i^{(O)}$, the response indicator variables for (*non-*)omissions, and $D_i^{(N)}$ the indicator of reached items. It is important to note that $\boldsymbol{D}$, $\boldsymbol{D^{(O)}}$, and $\boldsymbol{D^{(N)}}$ suffer itself from missing data if planned missing data exist due to not-administered items. If an item i was not presented, then it is unknown whether a test taker would have reached and answered the item or not. Hence $D_i^{(N)}$, $D_i^{(O)}$, and D_i, respectively, are missing. In all models discussed in this work, planned missing data due to not-administered items were assumed to be MCAR. This is reasonable in most real applications. In this case, the missing data mechanism w.r.t. $\boldsymbol{D}$ is MCAR as well. However, if the administration of booklets and items depends on covariates, such as pre-tests, type of schools, or other factors, then these variables need to be included since the missing data mechanism is then MAR given $\boldsymbol{Z}$. As outlined in Section 4.5.6, missing data in $\boldsymbol{D^{(O)}}$ result not only from not presented items but also from not-reached items. If a not-reached item would have been reached, then it is unknown whether it was answered or omitted. If the tendency to omit items depends on the number of not-reached items, then missingness in $\boldsymbol{D^{(O)}}$ is also nonignorable. In this case, an appropriate model for $\boldsymbol{D^{(O)}}$ or a suited function $f(\boldsymbol{D^{(O)}})$ needs to be found first. If a latent response propensity is modeled based on $\boldsymbol{D^{(O)}}$, then functions $f(\boldsymbol{D^{(N)}})$ should be included in the background model (LRM). In a next step, a joint model for omitted

and not-reached items can be used, that combines an MIRT model and a latent regression model. However, if the number of items Y_i as well as the number of latent dimensions ξ_m are large, then the estimates $\hat{\boldsymbol{\theta}}$ should be used together with $f(\boldsymbol{D}^{(N)})$ as independent variables in a latent regression $E[\boldsymbol{\xi} \mid \hat{\boldsymbol{\theta}}, f(\boldsymbol{D}^{(N)})]$. The complexity of both sub-models of $\boldsymbol{Y}$ and $\boldsymbol{D}$ can lead to a joint model with too many parameters, which is simply too complicated for application. Model complexity is a limiting factor especially for small sample. Unfortunately, there is scarcely any experience with all the proposed models for nonignorable missing responses, so that no clear recommendations can be given with respect to sample size requirements. LRMs and MG-IRT models for item nonresponses are more parsimonious than MIRT models for nonignorable missing data and might be preferred in moderate sample sizes. The model complexity can also be reduced by skipping all response indicators D_i from $\boldsymbol{D}$ that have no or very small proportions of missing responses. The item parameters of these indicators are difficult to estimate unless the sample size is very large. $\boldsymbol{D}$ can be partitioned in such cases, so that response indicators of items with substantial proportions of nonresponses are used as indicators in a measurement model of $\boldsymbol{\theta}$, whereas functions of the remaining indicators are used as independent variables in an additional LRM or as a grouping variable in multiple group MIRT models.

In order to find the best suited model and/or appropriate functions $f(\boldsymbol{D})$ or $f(\boldsymbol{D}^{(O)})$, it is strongly recommended to examine $\boldsymbol{D}$ by means of exploratory methods, such as item clustering (e. g. Reckase, 2009) or item factor analysis (e. g. Wirth & Edwards, 2007). The response indicators are not rationally constructed items, therefore the purely theoretical determination of the dimensionality of $\boldsymbol{\theta}$ is questionable. For example, M*plus* (Muthén & Muthén, 1998 - 2010) allows for exploratory factor analysis with dichotomous items based on tetrachoric correlations. Further methods for assessing the underlying dimensionality in the case of dichotomous items have been proposed (Jasper, 2010; Reckase, 2009; Roussos et al., 1998; Stout et al., 1996; Tate, 2003).

No exploratory or confirmatory factor analytical models should be utilized for $\boldsymbol{D}^{(N)}$, since the essential assumption of local stochastic independence is violated. If the items are answered in the same order the sum of reached items, then $S^{(N)}$ is always sufficient. All information of $\boldsymbol{D}^{(N)}$ is given by $S^{(N)}$. Of course, if the order of items varies, then $S^{(N)}$ does not preserve all information of $\boldsymbol{D}^{(N)}$ any more. If the information about the order of responded items is known, then $S^{(N)}$ can still be used. For example, if the item order depends on the booklet, then indicator variables I_h of the booklets $h = 1, \ldots, H$ can be included. Interactions between $S^{(N)}$ and the booklet indicators in a latent regression $E[(\boldsymbol{\xi}, \boldsymbol{\theta}) \mid S^{(N)}, I_1, \ldots, I_H]$ are appropriate to account for different item orders and/or

different sets of presented items in the booklets.

In general, the possibility of non-linear relations between $f(\boldsymbol{D})$, $f(\boldsymbol{D}^{(N)})$, $f(\boldsymbol{D}^{(O)})$ or $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ should be considered. If interactions and nonlinearities are expected, then LRMs can be superior to MIRT models that allow only for linear relations between the dimensions ξ_m and θ_l[4].

Limited software capabilities might also limit the range of applicable models. M*plus* is the only program that can estimate all models presented here. However, MG-IRT models and LRMs for nonignorable missing data are closely related. Hence, even if MIRT models or LRM cannot be applied in a particular software, MG-IRT models based on discrete functions $f(\boldsymbol{D})$ can considerably reduce the bias due to item nonresponses (Rose et al., 2010). Many MIRT software packages do not allow to specify complex nonlinear constraints with respect to item discrimination parameters. Additionally, bi-factor analysis is commonly used to reduce computational burden in MIRT modelling (Gibbons & Hedeker, 1992; Gibbons et al., 2007). In such cases, the relaxed 2PL-W$_{Res}$MIRT model can still be applied. This model is not equivalent to the 2PL-BMIRT model in terms of model fit, but yields unbiased item and person parameter if the model assumptions hold true. There might be other limitations in the available software. However, most IRT software packages allow at least for one of the model-based approaches discussed in this work: MIRT models, LRMs, MG-IRT models, or combinations of these approaches.

To sum up, the final missing data model should be established stepwise. If it can be assumed that the missing data mechanism is MCAR or MAR, then $\boldsymbol{D}$ needs *not* to be included in the model. If the missing data mechanism is MAR given $\boldsymbol{Z}$ or $(\boldsymbol{Y}, \boldsymbol{Z})$, then the covariate $\boldsymbol{Z}$ needs to be included in the model. If the nonresponse mechanism is suspected to be nonignorable, then $\boldsymbol{D}$ needs to be included in a joint model $(\boldsymbol{Y}, \boldsymbol{D})$. If both - omitted and not-reached items - needs to be considered, they need to be treated differently in a joint model including functions $f(\boldsymbol{D}^{(N)})$ and an appropriate model of $\boldsymbol{D}^{(O)}$.

5.3 Future Research

In this work, existing ad-hoc missing data methods were critically examined and existing model-based methods especially for item nonresponses have been extended. The initial examination of ad-hoc missing data methods, such as IAS and PCS, was motivated by their widespread use that contradicts the persistent criticism against their application.

[4] In this section, future research, the possibility of mixture MIRT models for latent interactions between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ will be discussed.

Its plausibility is tempting but nevertheless misleading. A closer analytical examination revealed the strong assumptions underlying these methods and the inconsistencies with stochastic IRT measurement models commonly applied in educational and psychological assessments. These results underlined that elaborate missing data methods are required. Data augmentation methods and model-based procedures have been proved promising in many applications when missing data needs to be taken into account. Data augmentation methods for item nonresponses in dichotomous items works well given the missing data mechanism is MAR. Appropriate model-based approaches have been developed for situations when the missing data is MAR or NMAR. Models for item nonresponses that are MAR were only briefly reviewed here. The focus was clearly on models for nonignorable item nonresponses. Since the late 1990s and the first decade of this new millennium, MIRT models have been proposed to handle missing responses that are NMAR. Here these models were related to existing models for missing data such as SLM and PMM. Furthermore, the relation between different existing between- and within- item multidimensional MIRT models were examined, and a common framework for these models was introduced. With these class of models, nonignorable missing data can be taken into account in many available software packages that do not allow for multidimensional IRT modeling.

However, there remain unsolved problems and unanswered questions that should be addressed in future research. In this work, only dichotomous items Y_i were considered. Many results and conclusions in this work cannot simply be generalized to polytomous items. It can be expected that the model-based approaches examined here work well with 1PL- and 2P-IRT models for ordinal items Y_i. In three-parameter models, parameter estimation is generally difficult even in absence of missing data but might become even more challenging due to nonignorable missing responses. The effect of item nonresponses and the inclusion of $\boldsymbol{D}$ in a joint model need to be investigated in future studies.

But even for the case of dichotomous items, there are still unanswered questions. All models discussed and developed here have restrictions reflecting certain assumptions, which can be questionable and not justifiable in some applications. For example, MIRT models rest upon the assumption of local stochastic independence of all manifest variables Y_i and D_i. Specifically, it is assumed that $D_i \perp (\boldsymbol{Y}, \boldsymbol{D}_{-i}) \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$. This implies $D_i \perp Y_i \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$. From the B-MIRT model follows that conditional stochastic independence $D_i \perp (Y_i, \boldsymbol{\xi}) \,|\, \boldsymbol{\theta}$ is assumed. In other words, the probability to respond to item i is independent of the item response (right or wrong) and the latent proficiency *given* the latent response propensity $\boldsymbol{\theta}$. In other words, in the MIRT models a latent variable un-

derlying $\boldsymbol{D}$ is constructed that completely explains all stochastic dependencies between each D_i and Y_i as well as each D_i and $\boldsymbol{\xi}$, respectively. Similarly, all pairwise stochastic dependencies between D_i and Y_i are implied by the latent covariance structure of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. Why are these assumptions critical? For example, if the persons tendency to respond to a particular item i depends on their subjective expectation of giving the (in)correct answer, then test takers tend more to omit an item if they expect to answer incorrectly, while they tend more to respond to an item if they feel to answer correctly. If it is further assumed that this subjective judge of correctness of the answer is not completely wrong, then local stochastic independence $D_i \perp Y_i \,|\, (\boldsymbol{\xi}, \boldsymbol{\theta})$ is violated. Even in this case it is expected that the MIRT models for nonignorable missing data will reduce the bias since information of D_i with respect to test performance is taken into account. However, the model is actually misspecified and potentially the bias may not be eliminated completely. Further research and simulation studies including conditional stochastic independence between items and response indicators are required. The robustness of the MIRT models, LRM, and MG-IRT models under local stochastic dependence needs to be investigated. Additionally, the development of less restrictive and more advanced models that allow for local stochastic dependencies would be an important step. As long as such models are not available, here it is argued that the latent variable model underlying $\boldsymbol{D}$ is as flexible as possible.

There are many more plausible models that might be worth to be considered in future research. For example, so far it was assumed the latent response propensity is a uni- or mulitdimensional continuous variable. Alternatively, it can be assumed that latent classes exist with typical missing data patterns. In this case, latent class models would be an appropriate choice to model $\boldsymbol{D}$. In fact, mixture modeling as implemented in Mplus (Muthén & Muthén, 1998 - 2010) allows for concurrent estimation of an LCA based on $\boldsymbol{D}$ and an IRT measurement model of a continuous uni- or multidimensional latent variable $\boldsymbol{\xi}$. Alternatively, mixture models that combine continuous latent response propensities and unobserved heterogeneity in $\boldsymbol{\theta}$ and the measurement model based on $\boldsymbol{D}$ might be a reasonable choice.

The MIRT models discussed here allow only for additive effects. However, interactions between latent variables with respect to the response indicators are thinkable. For example, the tendency to respond to item i depends on a general latent response propensity θ and the interaction between the latent ability ξ and θ. For example, let $P(D_i = 1 \,|\, \xi, \theta) = G[\gamma_{0i} + \gamma_{1i}\xi + (\gamma_{2i} - \gamma_{3i}\xi)\theta]$, with $G[]$ as the response function. In this case, the probability to answer to item i depends more strongly on the latent response propensity, the lower the ability is. IRT models that allow for interactions between latent

variables with respect to the manifest variables of the measurement models were recently introduced (Rizopoulos & Moustaki, 2008). However, apart from *ltm* (Rizopoulos, 2006) there is hardly any software that allow to fit such models. Insofar, the development of less restrictive models for nonignorable missing data depends also on the further development of IRT models and their implementation in available software.

Finally, it should be noted that data augmentation methods have also been discussed for missing data that are NMAR (Durrant & Skinner, 2006; Rubin, 1987). Multiple imputations with an imputation model that accounts for nonignorability of missingness would dispense with the need for joint model of $\boldsymbol{Y}$ and $\boldsymbol{D}$ in the estimation of the target measurement model. Multiple imputations for ignorable item nonresponses rest upon little assumptions with respect to the dimensional structure underlying $\boldsymbol{Y}$ and result in unbiased item and person parameter estimates even if the proportion of missing data is large (Van Buuren, 2010). If appropriate imputation models could be developed for nonignorable missing responses, then MI could become an interesting alternative to complex model-based approaches.

This work has broadened the range of models that are appropriate in many applications. Bias in item and person parameter estimates can be eliminated if the assumptions are met. Even if the assumption of local stochastic independence of the response indicators is violated, it is expected that the bias can be reduced. However, more research is required for a further development of missing data methods to handle situations in which existing approaches with their specific assumptions are inappropriate.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255–278. doi: 10.1207/s15324818ame0704_1

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51. doi: 10.1111/j.1745-3992.2003.tb00136.x

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172. doi: 10.1016/j.stueduc.2005.05.008

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. doi: 10.1177/0146621697211001

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NJ: Wiley. doi: 10.1002/0471249688

Allison, P. D. (2001). *Missing data*. Sage University Papers Series on Quantitative Applications in the Social Sience, 07-136. Thousend Oaks, CA: Sage.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 545–557. doi: 10.1037/0021-843X.112.4.545

Amelang, M., & Zielinski, W. (2001). *Psychologische Diagnostik und Intervention* (3. ed.). Springer: Berlin. doi: 10.1007/3-540-28507-5

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, *24*(1-2), 3–61. doi: 10.1016/0304-4076(84)90074-5

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco, CA: Jossey-Bass. doi: 10.2307/270846

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: CA: Sage.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (p. 243 - 277). Mahwah, NJ: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. O. (2010). *Weighted least squares estimation with missing data* (Mplus Webnote). Los Angeles, CA: Muthén & Muthén. Retrieved from www.statmodel.com\download\GstrucMissingRevision.pdf

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*(2), 111–141. doi: 10.1177/014662168701100201

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC. doi: 10.2307/2532822

Baraldi, A. N., & Enders, C. K. (2010, February 22). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37. doi: 10.1016/j.jsp.2009.10.001

Barndorff-Nielsen, O. (1976). Factorization of likelihood functions for full exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, *38*(1), 37–44.

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. doi: 10.1007/BF02293801

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT 4 [Computer software manual]. Chicago, IL: Scientific Software International.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. doi: 10.1177/014662168200600405

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

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634. doi: 10.1146/annurev.psych.53.100901.135239

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psy-*

chometrics. Cambridge: Cambridge University Press. Hardcover.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. doi: 10.1007/s11336-006-1447-6

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research & Perspective*, *6*(1), 25–53. doi: 10.1080/15366360802035497

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. doi: 10.1016/S0160-2896(02)00082-X

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. doi: 10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bowman, A., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-plus illustrations*. Oxford: Oxford University Press.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: ACE/Praeger Publishers.

Cai, L. (2010). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. doi: 10.3102/1076998609353115

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (p. 67-128). Chicago, IL: Rand-McNally.

Chen, H., & Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, *86*(1), 1–13. doi: 10.1093/biomet/86.1.1

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. doi: 10.1037/0022-3514.52.1.81

Cox, D. R., & Wermuth, N. (1999). Likelihood factorizations mixed discrete and continuous variables. *Scandinavian Journal of Statistics*, *26*, 209–222. doi: 10.1111/1467-9469.00145

Cramér, H. (1949). On the factorization of certain proability distributions. *Arkiv för*

Matematik, *1*(7), 61–65. doi: 10.1007/BF02590468

Cronbach, E. L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/BF02590468

Culbertson, M. (2011, April). *Is it wrong? Handling missing responses in IRT*. Speech presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NJ: Guilford Press.

de Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 213–234. doi: 10.1111/j.1745-3984.2001.tb01124.x

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, *33*(6), 465-485. doi:10.1177/0146621608329890. doi: 10.1177/0146621608329890

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. *Applied Measurement in Education*, *15*, 15–31. doi: 10.1207/S15324818AME1501_02

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing*, *7*, 247–252. doi: 10.1023/A:1018598421607

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. Series B (Methodological)*, *39*(1), 1–38. doi: 10.2307/2984875

do Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: SSI - Scientific Software International, Inc.

Dubin, J. A., & Rivers, D. (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods Research*, *18*(2-3), 360–390. doi: 10.1177/0049124189018002006

Durrant, G. B., & Skinner, C. (2006). Using data augmentation to correct for non-ignorable non-response when surrogate data are available: An application to the distribution of hourly pay. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 605–623. doi: 10.1111/j.1467-985X.2006.00398.x

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. doi:

10.1037//1082-989X.5.2.155

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, *65*(3), 457–483. doi: 10.1093/biomet/65.3.457

Eid, M., & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association. doi: 10.1037/11383-000

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*(3), 230–253. doi: 10.1037/a0013219

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515. doi: 10.1007/BF02294487

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2001a). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713-740. doi: 10.1177/0013164401615001

Enders, C. K. (2001b). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*(1), 128–141. doi: 10.1207/S15328007SEM0801_7

Enders, C. K. (2005). Maximum likelihood estimation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (p. 1164 - 1170). New York, NJ: Wiley.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Enders, C. K., & Bandalos, D. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430–457.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. doi: 10.1177/0013164498058003001

Finkelman, M. D., Hooker, G., & Wang, Z. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. *Journal of Educational and Behavioral Statistics*, *35*(6), 744–761. 10.3102/1076998610381402.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet

designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. doi: 10.1111/j.1745-3992.2009.00154.x

Frey, A., & Seitz, N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies In Educational Evaluation*, *35*(2-3), 89–94. doi: 10.1016/j.stueduc.2009.10.007

Geiser, C., Eid, M., Nussbeck, F., Courvoisier, D., & Cole, D. (2010). Multitrait-multimethod change modelling. *AStA Advances in Statistical Analysis*, *94*, 185–201. doi: 10.1007/s10182-010-0127-0

Geiser, C., & Lockhart, G. (2012, February 6). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, Advance online publication. doi: 10.1037/a0026977

Gelman, A. (2002). Posterior distribution. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics* (Vol. 1, pp. 1627–1628). John Wiley & Sons Inc. doi: 10.1002/9780470057339

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. London: Chapman & Hall.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., … Stover, A. (2007, 1). Full-Information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19. doi: 10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. (1992). Full-Information item bifactor analysis. *Psychometrika*, *57*(3), 423–436.

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational and Behavioral Statistics*, *13*(1), 45–52. doi: 10.3102/10769986013001045

Glas, C. A. W. (2006). *Violations of ignorability in computerized adaptive testing* (Research Report No. Report 04-04). Enschede, The Netherlands: University of Twente.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907–922. doi: 10.1177/0013164408315262

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonrespons. In H. Wainer (Ed.), *Drawing inferences from self-selected samples.* New York, NJ: Springer. doi: 10.1007/978-1-4612-4976-4_10

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10(1)*, 80–100. doi: 10.1207/S15328007SEM1001_4

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*(1), 549–576. doi: 10.1146/annurev.psych.58.110405.085530

Graham, J. W., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. doi: 10.1037/1082-989X.11.4.323

Green, W. H. (2012). *Econometric analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Guo, S., & Fraser, M. W. (2009). *Propensity score analysis: Statistical methods and applications*. Newsbury Park, CA: Sage.

Guttman, L. (1945). A basis for test-retest reliability. *Psychometrika*, *10*, 255–282. doi: 10.1007/BF02288892

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction.* Princeton, NJ: Princeton University Press.

Hambleton, R. K., & Jones, R. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newsbury Park, CA: Sage.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional irt models with within-item and between-item multidimensionality. *Journal of Psychology*, *216*(2), 89–101. doi: 10.1027/0044-3409.216.2.89

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2-3), 57–63.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, *5*, 475-492.

Heckman, J. (1979, January). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–61. doi: 10.2307/1912352

Heijmans, R. (1999). When does the expectation of a ratio equal the ratio of expectations?

Statistical Papers, *40*(1), 107–115. doi: 10.1007/BF02927114

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika*, *81*(4), 701–708. doi: 10.1093/biomet/81.4.701

Held, L. (2008). *Methoden der statistischen Inferenz*. Berlin: Spektrum Akademischer Verlag.

Hoijtink, H., & Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika*, *61*(2), 313–330. doi: 10.1007/BF02294342

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Hooker, G. (2010). On separable tests, correlated priors, and paradoxical results in multidimensional item response theory. *Psychometrika*, *75*, 694–707. doi: 10.1007/s11336-010-9181-5

Hooker, G., Finkelman, M. D., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, *74*, 419–442. doi: 10.1007/s11336-009-9111-6

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation*. Upper Saddle River, NJ: Prentice Hall.

Hsu, Y. (2000). On the Bock-Aitkin procedure - from an EM algorithm perspective. *Psychometrika*, *65*, 547-549. doi: 10.1007/BF02296345

Hu, B., Shao, J., & Palta, M. (2006). Pseudo-R^2 in logistic regression model. *Statistica Sinica*, *16*, 847-860.

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity*, *34*, 331-351. doi: 10.1023/A:1004782230065

Jasper, F. (2010). Applied dimensionality and test structure assessment with the START-M mathematics test. *The International Journal of Educational and Psychological Assessment*, *6*(1), 104 - 125.

Jeffrey, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement*, *29*(2), pp. 95-110. doi: 10.1111/j.1745-3984.1992.tb00369.x

Jöreskog, K. G., & Moustaki, I. (2000). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387. doi: 10.1207/S15327906347-387

Jöreskog, K. G., & Sörbom, D. (1997). *LISREL 8: User's reference guide*. Lincolnwood,

IL: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL for Windows [Computer software manual]. Lincolnwood, IL: Scientific Software International, Inc..

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 136–153. doi: 10.1080/10705510701758406

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, *13*, 236–247. doi: 10.1214/ss/1028905886

Kim, J. K., & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, *35*(4), 501–514. doi: 10.1002/cjs.5550350403

Koop, J. C. (1972). On the derivation of expected value and variance of ratios without the use of infinite series expansions. *Metrika*, *19*, 156-170. doi: 10.1007/BF01893291

Korobko, O. K., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*, 137–155. doi: 10.1111/j.1745-3984.2007.00057.x

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, U. K.: CUP.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden werten in der psychologischen forschung. *Psychologische Rundschau*, *58*, 103–117. doi: 10.1026/0033-3042.58.2.103

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics*, *34*(4), 433-463. doi: 10.3102/1076998609332757

Li, L., Shen, C., Li, X., & Robins, J. M. (2011). On weighting approaches for missing data. *Statistical Methods in Medical Research*, 1–17. doi: 10.1177/0962280211403597

Little, R. J. A. (1988a). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Little, R. J. A. (1988b). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 1198–1202. doi: 10.1080/01621459.1988.10478722

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125–134. doi: 10.2307/2290705

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112–1121. doi: 10.1080/01621459.1995.10476615

Little, R. J. A. (2008). Selection and pattern-mixture models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (p. 409-432). Chapman & Hall/CRC Press.

Little, R. J. A., & Rubin, D. B. (1984). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, *37*(3), 218–220. doi: 10.1080/00031305.1983.10483106

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: L. Erlbaum Associates.

Lord, F. M. (1983a). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477–482. doi: 10.1007/BF02293689

Lord, F. M. (1983b). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233–245. doi: 10.1007/BF02294018

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading Mass: Addison-Welsley.

Lubke, G. H., & Muthén, B. O. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, *Vol. 11*, 514–534. doi: 10.1207/s15328007sem1104_2

Ludlow, L. H., & O'Leary, M. (1999). Scoring Omitted and Not-Reached Items: Practical Data Analysis Implications. *Educational and Psychological Measurement*, *59*(4), 615–630. doi: 10.1177/0013164499594004

McCaffrey, D., & Lockwood, J. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, *5*(2A), 773–797. doi: 10.1214/10-AOAS405

McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figuerdo, A. J. (2007). *Missing Data:*

A Gentle Introduction. New York, NJ: Guilforg Press.

McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions* (Vol. 382). Hoboken, NJ: John Wiley and Sons. doi: 10.1002/9780470191613

Meiser, T. (2007). Rasch model for longitudinal data. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution rasch models.* New York, NJ: Springer. doi: 10.1007/978-0-387-49839-3_12

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 121–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance , factor analysis, and factorial invariance. *Psychometrika*, *58*, 521–543. doi: 10.1007/BF02294825

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational and Behavioral Statistics*, *11*(1), 3–31. doi: 10.3102/10769986011001003

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*(1), 81–91. doi: 10 .1177/014662168701100106

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of rasch item difficulty parameters. *Applied Psychological Measurement*, *12*(3), 281–296. doi: 10.1177/014662168801200306

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161. doi: 10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30). Princeton, NY: Educational Testing Service.

Moosbrugger, H., & Kelava, A. (2011). *Testtheorie und fragebogenkonstruktion*. Springer.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, *163*, 445–459. doi: 10.1111/1467-985X.00177

Moustaki, I., & O'Muircheartaigh, C. (2000). A one dimensional latent trait model to infer attitude from nonresponse for nominal data. *Statistica*, *10*, 259-276.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. doi: 10.1177/ 014662169201600206

Muraki, E., & Bock, R. D. (2002). PARSCALE: IRT based test scoring and item analysis for graded items and rating scales (Version 4) [Computer software manual]. Chicago, IL: Scientific Software International, Inc.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560. doi: 10.1007/BF02293813

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49*, 115 - 132. doi:10.1007/BF02294210.

Muthén, B. O. (1998 - 2004). *Mplus Technical Appendices* (Technical Report). Los Angelas, CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/download/techappen.pdf

Muthén, B. O., do Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* (Accepted for Publication in Psychometrika)

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *42*, 431–462. doi: 10.1007/BF02294365

Muthén, B. O., & Muthén, L. K. (1998 - 2010). Mplus User´s Guide (Version 6) [Computer software manual]. Los Angeles, CA: Muthén and Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–693. doi: 10.1093/biomet/78.3.691

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitudes scales. *Journal of Royal Statistic Society*, *162*, 177–194. doi: 10.1111/1467-985X.00129

Organisation for Economic Co-operation and Development. (2009a). *Pisa 2006 data analysis manual* (2nd ed.). Paris: Author.

Organisation for Economic Co-operation and Development. (2009b). *Pisa 2006 technical report*. Paris: Author.

Pauler, D. K., McCoy, S., & Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine*, *22*(5), 795–809. doi: 10.1002/sim.1397

Peugh, J. L., & Enders, C. K. (Winter 2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556. doi:10.3102/00346543074004525. doi: 10.3102/00346543074004525

Pohl, S., Gräfe, & Hardt, K. (2011, September). Ignorability & Modellierung von fehlenden Werten in Kompetenztests. *Speech presented at 10. Tagung der Fachgruppe Methoden & Evaluation der DGPs*, *Bamberg*, Germany.

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research*, *45*(1), 45–72. doi: 10.1080/00273170903504729

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, *14*(1), 53–68. doi: 10.1111/1467-6419.00104

Quesenberry, C. P., & Jewell, N. P. (1986). Regression analysis based on stratified samples. *Biometrika*, *73*(3), 605–614. doi: 10.1093/biomet/73.3.605

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Raghunathan, T. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, *25*, 99–117. doi: 10.1146/annurev.publhealth.25.102802.124410

Raghunathan, T., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*, 85-95.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*(429), 54–63. doi: 10.1080/01621459.1995.10476488

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630. doi: 10.1007/BF02294494

Rasch, G. (1960). *Probalistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research (reprinted 1980 by University of Chicago Press: Chicago).

Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(1), 142–149. doi: 10.1207/S15328007SEM0801_8

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research*, *34*(2), 199–244. doi: 10.1207/S15327906Mb340204

Reckase, M. D. (1985). The difficulty of tests that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412. doi: 10.1177/014662168500900409

Reckase, M. D. (1997). Handbook of modern IRT. In W. J. van der Linden & R. K. Ham-

bleton (Eds.), (p. 271-286). Springer.

Reckase, M. D. (2009). *Multidimensional item response theory*. London, England: Springer. doi:10.1007/978-0-387-89976-3.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *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*(2), 415–438. doi: 10.1348/000711007X213963

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with IRT* (Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Huber.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*(1), 1–30. doi: 10.1111/j.1745-3984.1998.tb00525.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63(3)*, 581-592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponses in surveys*. New York, NJ: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. doi: 10.1080/01621459.1996.10476908

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, *28*(9), 1420–1423. doi: 10.1002/sim.3565

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, No. 17*, 100–113.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, England: Chapman & Hall. doi: 10.1201/9781439821862

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi: 10.1037/1082-989X.7.2.147

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. doi: 10.1007/BF02294343

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–74). Dordrecht/Boston/London: Kluwer.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and

differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–230). Hillsdale, N.J.: Lawrence Erlbaum.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9780203489437

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*, 309-331. doi: 10.1207/s15327906mbr2103_3

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60.

Steyer, R. (2001). Classical test theory. In C. Ragin & T. Cook (Eds.), *International encyclopedia of the social and behavioural sciences. Logic of inquiry and research design* (pp. 481–520). Pergamon, Oxford.: Oxford University Press.

Steyer, R. (2002). *Wahrscheinlichkeit und Regression*. Berlin: Springer.

Steyer, R., & Eid, M. (2001). *Messen und Testen*. Berlin: Springer. doi: 10.1007/978-3-642-56924-1

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2*(1), 21–33.

Steyer, R., Krambeer, S., & Hannover, W. (2004). Modeling latent trait-change. In K. Van Montfort & J. Oud (Eds.), *Recent developments on structural equation models: Theory and applications* (Vol. 19, pp. 337–357). Dordrecht, Netherlands: Kluwer Academic Publishers.

Steyer, R., Nagel, W., Partchev, I., & Mayer, A. (in press). *Probability and regression*. New York, NJ: Springer.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. doi: 10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331–354. doi: 10.1177/014662169602000403

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi:

10.1007/BF02294363

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*(3), 159–203. doi: 10.1177/0146621603027003001

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Thissen, D. M. (1976). Information in wrong responses to the raven progressive matrices. *Journal of Educational Measurement*, *13*(3), 201–214. doi: 10.1111/j.1745-3984.1976.tb00011.x

Thissen, D. M., Chen, W. H., & Bock, R. D. (2003). MULTILOG (Version 7) [Computer software manual]. Lincolnwood, IL: Scientific Software International, Inc.

Thissen, D. M., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. doi: 10.1007/BF02295596

Thomas, N., Raghunathan, T., Schenker, N., Katzoff, M., & Johnson, C. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*(2), 217.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Toomet, O., & Henningsen, A. (2008, 7 29). Sample selection models in R: Package sample selection. *Journal of Statistical Software*, *27*(7), 1-23.

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York (NJ): Springer Verlag.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specifcation. *Statistical Methods in Medical Research*, *16*, 219–242. doi: 10.1177/0962280206074463

Van Buuren, S. (2010). Item imputation without specifiying scale structure. *Methodology - European Journal of Research Methods in the Behavioral and Social Sciences*, *6*, 31–36. doi: 10.1027/1614-2241/a000004

Van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, *28*(5), 317–331. doi: 10.1177/0146621604264870

Verhelst, N. D., Glas, C. A. W., & De Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York, NJ: Springer.

von Davier, M. (2005). *A general diagnostic model applied to language testing data*

(Research Report No. RR-05-16). Princeton, NJ: ETS.

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151–176). Cambridge, MA: Hogrefe & Huber.

von Davier, M., Xu, X., & Carstensen, C. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*(2), 318–336. doi:dx.doi.org/10.1007/s11336-011-9202-z. doi: dx.doi.org/10.1007/s11336-011-9202-z

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the national assessment of educational progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics.* Amsterdam: Elsevier.

Wang, W. C., Wilson, M., & Adams, R. (1997). Rasch models for multidimensionality between and within items. In M. Wilson, G. J. Englehard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4). Greenwich, CT,London: Ablex Publishing Corporation.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. doi: 10.1007/BF02294627

Winship, C., & Mare, R. (1992). Models for sample selection bias. *Annual Review of Sociology*, 327–350. doi: 10.1146/annurev.so.18.080192.001551

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. doi: 10.1037/1082-989X.12.1.58

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*(2), 1281–1301. doi: 10.1016/j.jeconom.2007.02.002

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ACER ConQest: Generalised Item Response Modelling Software [Computer software manual]. Mebourne: Australia.

Zieky, M. J. (2006). Fairness reviews in assessment. In S. Downing & T. Haladyna (Eds.), *In handbook of test development.* Mahwah, NJ: Lawrence Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items (Version 3) [Computer software manual]. Chicago, IL: Scientific Software International, Inc.

Appendix

Appendix A

In Chapter 3 the impact of missing data on item and person parameter was studied using a simulation study. Details of the simulation study can be found in Chapter 3 (see page 46). In this Appendix the results of the simulation study are presented in detail. Table 5.2 shows the mean bias $Bias(\lambda)$, as defined in Equation 3.4, of item parameter estimates $\hat{\beta}_i$ and $\hat{\alpha}_i$ as well as the ML-, WML-, and EAP person parameter estimates.

Table 5.2: Mean bias of estimated item difficulties $\hat{\beta}_i$ and item discriminations $\hat{\alpha}_i$ and person parameter estimates $\hat{\xi}_{ML}$, $\hat{\xi}_{WML}$, and $\hat{\xi}_{EAP}$.

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	11	0	-2	0	-0.001	0.036	0.011	-0.003	0
500	11	0	-2	0.25	-0.003	0.046	0.017	-0.007	-0.006
500	11	0	-2	0.5	0.007	0.027	0.024	-0.003	0.005
500	11	0	-2	0.8	0.009	0.037	0.022	-0.006	0
500	11	0	-1.5	0	0.001	0.032	0.006	-0.008	-0.004
500	11	0	-1.5	0.25	0.005	0.055	0.018	-0.008	-0.004
500	11	0	-1.5	0.5	0	0.046	0.017	-0.015	0
500	11	0	-1.5	0.8	0.009	0.041	0.02	-0.014	0.001
500	11	0	-1	0	-0.002	0.05	0.004	-0.01	-0.006
500	11	0	-1	0.25	0.002	0.029	0.025	-0.004	0.001
500	11	0	-1	0.5	0.006	0.072	0.004	-0.033	-0.004
500	11	0	-1	0.8	0.003	0.049	0.017	-0.025	-0.002
500	11	0	-0.5	0	-0.011	0.038	-0.006	-0.021	-0.007
500	11	0	-0.5	0.25	-0.001	0.054	0.029	-0.002	0.01
500	11	0	-0.5	0.5	0.01	0.071	0.003	-0.036	0.012
500	11	0	-0.5	0.8	0.007	0.106	0.004	-0.044	0
500	11	0	0	0	0.003	0.073	-0.006	-0.018	-0.002
500	11	0	0	0.25	-0.01	0.125	0.013	-0.019	0.001
500	11	0	0	0.5	-0.002	0.112	-0.035	-0.077	0
500	11	0	0	0.8	0.014	0.033	-0.01	-0.059	0.01
500	11	0.3	-2	0	-0.023	0.009	-0.005	-0.018	-0.003
500	11	0.3	-2	0.25	-0.01	0.034	0.013	-0.01	0.006
500	11	0.3	-2	0.5	-0.044	0.032	-0.01	-0.039	-0.012

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	11	0.3	-2	0.8	-0.012	0.003	0.006	-0.021	-0.002
500	11	0.3	-1.5	0	-0.036	0.02	-0.016	-0.03	-0.002
500	11	0.3	-1.5	0.25	-0.04	0.052	-0.007	-0.034	-0.01
500	11	0.3	-1.5	0.5	-0.027	0.029	-0.008	-0.039	-0.002
500	11	0.3	-1.5	0.8	-0.025	0.013	0.001	-0.033	-0.002
500	11	0.3	-1	0	-0.054	0.047	-0.025	-0.042	-0.005
500	11	0.3	-1	0.25	-0.035	0.049	-0.01	-0.041	-0.011
500	11	0.3	-1	0.5	-0.039	0.051	-0.011	-0.05	0.003
500	11	0.3	-1	0.8	-0.042	0.022	-0.004	-0.047	-0.001
500	11	0.3	-0.5	0	-0.05	0.056	-0.027	-0.043	0.006
500	11	0.3	-0.5	0.25	-0.032	0.054	0.002	-0.03	0.009
500	11	0.3	-0.5	0.5	-0.054	0.096	-0.032	-0.077	0.005
500	11	0.3	-0.5	0.8	-0.05	0.054	-0.024	-0.075	-0.004
500	11	0.3	0	0	-0.078	0.098	-0.036	-0.057	0.009
500	11	0.3	0	0.25	-0.077	0.059	-0.012	-0.05	0.004
500	11	0.3	0	0.5	-0.062	0.382	-0.062	-0.11	0.007
500	11	0.3	0	0.8	-0.065	0.088	-0.057	-0.113	-0.004
500	11	0.5	-2	0	-0.031	0.021	-0.021	-0.032	-0.003
500	11	0.5	-2	0.25	-0.021	-0.001	0.006	-0.016	0.01
500	11	0.5	-2	0.5	-0.032	0.007	-0.01	-0.035	0.003
500	11	0.5	-2	0.8	-0.037	0.013	-0.013	-0.041	-0.012
500	11	0.5	-1.5	0	-0.059	0.002	-0.036	-0.05	-0.008
500	11	0.5	-1.5	0.25	-0.048	0.018	-0.019	-0.046	-0.011
500	11	0.5	-1.5	0.5	-0.063	0.004	-0.019	-0.052	0.001
500	11	0.5	-1.5	0.8	-0.037	0.003	-0.01	-0.042	0
500	11	0.5	-1	0	-0.045	0.037	-0.026	-0.041	0.014
500	11	0.5	-1	0.25	-0.056	0.025	-0.014	-0.044	0.003
500	11	0.5	-1	0.5	-0.075	0.036	-0.039	-0.079	-0.004
500	11	0.5	-1	0.8	-0.05	0.052	-0.018	-0.06	0.004
500	11	0.5	-0.5	0	-0.088	0.029	-0.041	-0.061	0.013
500	11	0.5	-0.5	0.25	-0.072	0.035	-0.018	-0.053	0.008
500	11	0.5	-0.5	0.5	-0.092	0.073	-0.048	-0.096	0.005
500	11	0.5	-0.5	0.8	-0.075	0.037	-0.042	-0.093	-0.003
500	11	0.5	0	0	-0.137	0.078	-0.077	-0.104	0.003
500	11	0.5	0	0.25	-0.103	0.051	-0.033	-0.072	0.002
500	11	0.5	0	0.5	-0.144	0.166	-0.094	-0.148	-0.003
500	11	0.5	0	0.8	-0.106	0.922	-0.07	-0.13	0.004
500	11	0.8	-2	0	-0.052	0.012	-0.03	-0.041	0.006
500	11	0.8	-2	0.25	-0.051	0.009	-0.018	-0.04	0.002
500	11	0.8	-2	0.5	-0.066	0	-0.038	-0.062	-0.006

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	11	0.8	-2	0.8	-0.041	0.008	-0.018	-0.042	0
500	11	0.8	-1.5	0	-0.094	0	-0.056	-0.07	-0.006
500	11	0.8	-1.5	0.25	-0.071	-0.003	-0.022	-0.048	0.005
500	11	0.8	-1.5	0.5	-0.105	0.007	-0.057	-0.089	-0.01
500	11	0.8	-1.5	0.8	-0.062	-0.024	-0.031	-0.062	-0.001
500	11	0.8	-1	0	-0.112	-0.034	-0.067	-0.083	0.005
500	11	0.8	-1	0.25	-0.107	-0.021	-0.055	-0.087	-0.013
500	11	0.8	-1	0.5	-0.116	0.005	-0.056	-0.097	0.005
500	11	0.8	-1	0.8	-0.099	0.025	-0.047	-0.089	-0.005
500	11	0.8	-0.5	0	-0.153	0.02	-0.092	-0.117	0.006
500	11	0.8	-0.5	0.25	-0.113	0.013	-0.042	-0.077	0.016
500	11	0.8	-0.5	0.5	-0.157	0.035	-0.078	-0.131	0.009
500	11	0.8	-0.5	0.8	-0.139	-0.026	-0.067	-0.123	-0.003
500	11	0.8	0	0	-0.211	0.121	-0.127	-0.161	0.004
500	11	0.8	0	0.25	-0.18	0.205	-0.076	-0.121	-0.007
500	11	0.8	0	0.5	-0.221	0.063	-0.121	-0.185	0.006
500	11	0.8	0	0.8	-0.168	0.011	-0.098	-0.164	0.011
500	22	0	-2	0	0.01	0.013	0.014	0.01	0.009
500	22	0	-2	0.25	-0.004	0.009	0.009	0	-0.001
500	22	0	-2	0.5	0.013	0.012	0.021	0.01	0.012
500	22	0	-2	0.8	-0.006	0.024	0.011	-0.004	-0.003
500	22	0	-1.5	0	-0.005	0.04	-0.001	-0.006	-0.005
500	22	0	-1.5	0.25	0.005	0.026	0.013	0.003	0.004
500	22	0	-1.5	0.5	-0.008	0.029	0.012	-0.003	-0.002
500	22	0	-1.5	0.8	-0.003	0.022	0.017	-0.003	-0.002
500	22	0	-1	0	0.002	0.035	0.013	0.007	0.007
500	22	0	-1	0.25	0	0.026	0.014	0	0.001
500	22	0	-1	0.5	0.007	0.018	0.017	-0.002	0.001
500	22	0	-1	0.8	0.006	0.037	0.025	-0.001	0.003
500	22	0	-0.5	0	0.005	0.037	0.007	0.001	0
500	22	0	-0.5	0.25	-0.004	0.04	0.017	0	-0.002
500	22	0	-0.5	0.5	0.01	0.067	0.019	-0.006	-0.001
500	22	0	-0.5	0.8	0.01	0.036	0.025	-0.01	0.002
500	22	0	0	0	0.001	0.047	0.007	0.001	0.002
500	22	0	0	0.25	0.005	0.063	0.019	-0.002	-0.002
500	22	0	0	0.5	0.007	0.084	0.019	-0.013	0.007
500	22	0	0	0.8	-0.005	0.038	0.022	-0.024	0.001
500	22	0.3	-2	0	-0.006	0.007	-0.004	-0.006	0.005
500	22	0.3	-2	0.25	-0.019	0.012	-0.015	-0.022	-0.013
500	22	0.3	-2	0.5	-0.018	0.031	-0.009	-0.018	-0.004

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	22	0.3	-2	0.8	-0.012	0.019	0.002	-0.009	0.003
500	22	0.3	-1.5	0	-0.018	0.011	-0.017	-0.018	-0.002
500	22	0.3	-1.5	0.25	-0.015	0.011	-0.005	-0.012	0.002
500	22	0.3	-1.5	0.5	-0.018	0.037	-0.011	-0.022	-0.003
500	22	0.3	-1.5	0.8	-0.006	0.017	0	-0.015	0.003
500	22	0.3	-1	0	-0.028	0.007	-0.027	-0.028	-0.004
500	22	0.3	-1	0.25	-0.015	0.019	-0.01	-0.018	0.003
500	22	0.3	-1	0.5	-0.035	0.024	-0.014	-0.03	-0.004
500	22	0.3	-1	0.8	-0.032	0.032	-0.005	-0.027	0
500	22	0.3	-0.5	0	-0.044	0.041	-0.027	-0.029	0.008
500	22	0.3	-0.5	0.25	-0.042	0.02	-0.015	-0.027	0
500	22	0.3	-0.5	0.5	-0.04	0.043	-0.017	-0.04	0
500	22	0.3	-0.5	0.8	-0.052	0.041	-0.014	-0.047	-0.007
500	22	0.3	0	0	-0.06	0.017	-0.039	-0.044	-0.001
500	22	0.3	0	0.25	-0.045	0.059	-0.02	-0.036	0.001
500	22	0.3	0	0.5	-0.06	0.035	-0.024	-0.054	0.002
500	22	0.3	0	0.8	-0.033	0.04	-0.011	-0.052	0.01
500	22	0.5	-2	0	-0.012	-0.006	-0.019	-0.019	0
500	22	0.5	-2	0.25	-0.022	0.025	-0.013	-0.018	0
500	22	0.5	-2	0.5	-0.027	0.028	-0.021	-0.028	-0.006
500	22	0.5	-2	0.8	-0.03	0.023	-0.016	-0.026	-0.007
500	22	0.5	-1.5	0	-0.025	0.011	-0.023	-0.021	0.008
500	22	0.5	-1.5	0.25	-0.031	0.003	-0.021	-0.026	-0.002
500	22	0.5	-1.5	0.5	-0.018	0.027	-0.012	-0.019	0.014
500	22	0.5	-1.5	0.8	-0.028	0.027	-0.011	-0.024	0.005
500	22	0.5	-1	0	-0.05	-0.011	-0.046	-0.044	-0.002
500	22	0.5	-1	0.25	-0.036	0.026	-0.029	-0.035	-0.002
500	22	0.5	-1	0.5	-0.056	0.022	-0.028	-0.041	0.003
500	22	0.5	-1	0.8	-0.055	0.025	-0.031	-0.051	-0.009
500	22	0.5	-0.5	0	-0.064	0.003	-0.057	-0.056	0.002
500	22	0.5	-0.5	0.25	-0.067	0.04	-0.042	-0.051	-0.005
500	22	0.5	-0.5	0.5	-0.073	0.048	-0.043	-0.062	0.002
500	22	0.5	-0.5	0.8	-0.066	0.045	-0.029	-0.056	0.005
500	22	0.5	0	0	-0.093	0.027	-0.067	-0.069	0.003
500	22	0.5	0	0.25	-0.087	0.038	-0.044	-0.057	0.004
500	22	0.5	0	0.5	-0.104	0.036	-0.055	-0.084	0.002
500	22	0.5	0	0.8	-0.09	0.032	-0.047	-0.087	0
500	22	0.8	-2	0	-0.038	-0.004	-0.035	-0.032	-0.003
500	22	0.8	-2	0.25	-0.033	0	-0.03	-0.031	-0.004
500	22	0.8	-2	0.5	-0.031	0.009	-0.028	-0.031	0.003

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	22	0.8	-2	0.8	-0.043	0.012	-0.033	-0.039	-0.009
500	22	0.8	-1.5	0	-0.059	-0.012	-0.057	-0.053	-0.006
500	22	0.8	-1.5	0.25	-0.053	-0.009	-0.04	-0.043	-0.005
500	22	0.8	-1.5	0.5	-0.065	-0.01	-0.056	-0.062	-0.012
500	22	0.8	-1.5	0.8	-0.048	0.001	-0.034	-0.042	0.003
500	22	0.8	-1	0	-0.075	-0.011	-0.068	-0.061	0.007
500	22	0.8	-1	0.25	-0.065	-0.004	-0.054	-0.055	0.001
500	22	0.8	-1	0.5	-0.094	-0.01	-0.067	-0.077	-0.008
500	22	0.8	-1	0.8	-0.094	-0.015	-0.063	-0.078	-0.01
500	22	0.8	-0.5	0	-0.116	-0.034	-0.1	-0.096	-0.003
500	22	0.8	-0.5	0.25	-0.118	-0.006	-0.081	-0.086	-0.013
500	22	0.8	-0.5	0.5	-0.119	-0.014	-0.076	-0.091	0.007
500	22	0.8	-0.5	0.8	-0.108	-0.002	-0.07	-0.092	0
500	22	0.8	0	0	-0.169	-0.015	-0.125	-0.126	-0.002
500	22	0.8	0	0.25	-0.146	0.004	-0.092	-0.102	-0.001
500	22	0.8	0	0.5	-0.151	-0.016	-0.101	-0.127	0.002
500	22	0.8	0	0.8	-0.151	-0.024	-0.097	-0.134	-0.005
500	33	0	-2	0	-0.005	0.005	-0.003	-0.004	-0.006
500	33	0	-2	0.25	-0.002	0.024	-0.002	-0.005	-0.007
500	33	0	-2	0.5	-0.004	0.017	0.004	-0.003	-0.004
500	33	0	-2	0.8	0.009	0.009	0.011	0.003	0.002
500	33	0	-1.5	0	0.001	0.026	0	0	0
500	33	0	-1.5	0.25	0.005	0.012	0.006	0.002	0.001
500	33	0	-1.5	0.5	-0.004	0.022	0.006	-0.004	-0.004
500	33	0	-1.5	0.8	0.001	0.024	0.01	-0.002	-0.001
500	33	0	-1	0	-0.003	0.02	0	-0.001	-0.001
500	33	0	-1	0.25	0.008	0.025	0.016	0.01	0.01
500	33	0	-1	0.5	0.013	0.025	0.024	0.011	0.011
500	33	0	-1	0.8	-0.003	0.016	0.014	-0.004	-0.003
500	33	0	-0.5	0	0.001	0.051	-0.005	-0.006	-0.007
500	33	0	-0.5	0.25	-0.001	0.031	0.002	-0.007	-0.004
500	33	0	-0.5	0.5	0.004	0.062	0.02	0	0.001
500	33	0	-0.5	0.8	-0.002	0.042	0.021	-0.007	-0.001
500	33	0	0	0	-0.007	0.039	-0.005	-0.006	-0.005
500	33	0	0	0.25	0.002	0.053	0.004	-0.008	-0.003
500	33	0	0	0.5	0	0.039	0.016	-0.01	-0.003
500	33	0	0	0.8	0.001	0.075	0.034	-0.004	0.006
500	33	0.3	-2	0	-0.01	0.012	-0.013	-0.011	-0.003
500	33	0.3	-2	0.25	-0.002	0.027	-0.004	-0.004	0.007
500	33	0.3	-2	0.5	-0.004	0.025	0.003	-0.001	0.008

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	33	0.3	-2	0.8	-0.015	0.018	-0.008	-0.013	-0.005
500	33	0.3	-1.5	0	-0.003	0.02	-0.004	-0.002	0.012
500	33	0.3	-1.5	0.25	-0.026	0.025	-0.03	-0.03	-0.015
500	33	0.3	-1.5	0.5	-0.008	0.017	0	-0.006	0.009
500	33	0.3	-1.5	0.8	-0.01	0.024	-0.002	-0.011	0.003
500	33	0.3	-1	0	-0.014	0	-0.02	-0.016	0.005
500	33	0.3	-1	0.25	-0.031	0.032	-0.021	-0.024	-0.005
500	33	0.3	-1	0.5	-0.025	0.026	-0.007	-0.016	0.005
500	33	0.3	-1	0.8	-0.032	0.025	-0.014	-0.028	-0.008
500	33	0.3	-0.5	0	-0.03	0.024	-0.033	-0.029	0
500	33	0.3	-0.5	0.25	-0.029	0.037	-0.018	-0.022	0.006
500	33	0.3	-0.5	0.5	-0.035	0.034	-0.021	-0.034	-0.002
500	33	0.3	-0.5	0.8	-0.038	0.016	-0.015	-0.037	-0.007
500	33	0.3	0	0	-0.049	0.019	-0.041	-0.038	0
500	33	0.3	0	0.25	-0.046	0.034	-0.024	-0.032	0.006
500	33	0.3	0	0.5	-0.059	0.07	-0.026	-0.047	-0.003
500	33	0.3	0	0.8	-0.056	0.036	-0.021	-0.055	-0.011
500	33	0.5	-2	0	-0.021	0.007	-0.022	-0.019	-0.005
500	33	0.5	-2	0.25	-0.025	0.021	-0.021	-0.021	-0.005
500	33	0.5	-2	0.5	-0.009	0.025	-0.012	-0.014	0.002
500	33	0.5	-2	0.8	-0.016	0.009	-0.013	-0.017	-0.003
500	33	0.5	-1.5	0	-0.013	0.034	-0.021	-0.017	0.007
500	33	0.5	-1.5	0.25	-0.026	0.014	-0.024	-0.023	0.001
500	33	0.5	-1.5	0.5	-0.038	0.012	-0.026	-0.03	-0.005
500	33	0.5	-1.5	0.8	-0.03	0.012	-0.018	-0.024	-0.003
500	33	0.5	-1	0	-0.047	0.005	-0.042	-0.035	-0.001
500	33	0.5	-1	0.25	-0.034	0.012	-0.024	-0.024	0.009
500	33	0.5	-1	0.5	-0.038	0.022	-0.03	-0.036	-0.001
500	33	0.5	-1	0.8	-0.026	0.017	-0.007	-0.018	0.015
500	33	0.5	-0.5	0	-0.055	0.001	-0.054	-0.047	0
500	33	0.5	-0.5	0.25	-0.053	0.038	-0.045	-0.045	0.004
500	33	0.5	-0.5	0.5	-0.074	0.015	-0.047	-0.057	-0.008
500	33	0.5	-0.5	0.8	-0.04	0.022	-0.016	-0.034	0.012
500	33	0.5	0	0	-0.068	0.022	-0.063	-0.056	0.009
500	33	0.5	0	0.25	-0.076	0.05	-0.059	-0.064	0.001
500	33	0.5	0	0.5	-0.073	0.033	-0.043	-0.059	0.009
500	33	0.5	0	0.8	-0.072	0.039	-0.038	-0.066	0.002
500	33	0.8	-2	0	-0.035	0.002	-0.042	-0.038	-0.014
500	33	0.8	-2	0.25	-0.038	0.016	-0.041	-0.039	-0.013
500	33	0.8	-2	0.5	-0.025	-0.008	-0.022	-0.021	0.005

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
500	33	0.8	-2	0.8	-0.03	0.005	-0.027	-0.029	-0.006
500	33	0.8	-1.5	0	-0.042	-0.006	-0.053	-0.046	-0.008
500	33	0.8	-1.5	0.25	-0.038	-0.009	-0.038	-0.034	0.005
500	33	0.8	-1.5	0.5	-0.047	0.011	-0.039	-0.039	0.001
500	33	0.8	-1.5	0.8	-0.027	0.002	-0.025	-0.028	0.006
500	33	0.8	-1	0	-0.048	0.005	-0.049	-0.039	0.015
500	33	0.8	-1	0.25	-0.072	-0.011	-0.065	-0.061	-0.007
500	33	0.8	-1	0.5	-0.067	0.001	-0.06	-0.06	-0.002
500	33	0.8	-1	0.8	-0.036	-0.016	-0.032	-0.037	0.015
500	33	0.8	-0.5	0	-0.088	-0.02	-0.098	-0.086	-0.008
500	33	0.8	-0.5	0.25	-0.1	-0.018	-0.086	-0.083	-0.002
500	33	0.8	-0.5	0.5	-0.095	-0.011	-0.083	-0.087	-0.006
500	33	0.8	-0.5	0.8	-0.08	0.007	-0.06	-0.071	0.003
500	33	0.8	0	0	-0.126	-0.028	-0.118	-0.104	0.005
500	33	0.8	0	0.25	-0.141	-0.023	-0.116	-0.117	-0.01
500	33	0.8	0	0.5	-0.14	-0.018	-0.096	-0.106	0.005
500	33	0.8	0	0.8	-0.106	-0.005	-0.075	-0.095	0.011
1000	11	0	-2	0	-0.002	0.006	0.016	0.003	0.005
1000	11	0	-2	0.25	0.002	0.016	0.023	0	0.003
1000	11	0	-2	0.5	0.021	0.015	0.029	0.003	0.012
1000	11	0	-2	0.8	0.015	0.014	0.031	0.003	0.01
1000	11	0	-1.5	0	-0.011	0.017	-0.001	-0.016	-0.011
1000	11	0	-1.5	0.25	0.007	0.012	0.024	-0.002	0.001
1000	11	0	-1.5	0.5	0	0.014	0.014	-0.018	-0.002
1000	11	0	-1.5	0.8	-0.002	0.025	0.026	-0.009	0.003
1000	11	0	-1	0	0.005	0.01	0.009	-0.005	0.003
1000	11	0	-1	0.25	0.004	0.019	0.028	0	0.005
1000	11	0	-1	0.5	0.001	0.029	0.005	-0.033	-0.004
1000	11	0	-1	0.8	-0.007	0.015	0.013	-0.03	-0.005
1000	11	0	-0.5	0	-0.001	0.049	0.001	-0.013	0.001
1000	11	0	-0.5	0.25	-0.001	0.01	0.018	-0.014	0
1000	11	0	-0.5	0.5	0.002	0.031	-0.003	-0.044	0.005
1000	11	0	-0.5	0.8	-0.001	0.014	0.01	-0.037	0.003
1000	11	0	0	0	0.003	0.071	-0.003	-0.015	0.003
1000	11	0	0	0.25	0.005	0.044	0.016	-0.016	0.002
1000	11	0	0	0.5	0.003	0.06	-0.041	-0.082	-0.006
1000	11	0	0	0.8	0.003	0.046	-0.021	-0.072	-0.001
1000	11	0.3	-2	0	-0.018	0.003	-0.006	-0.018	-0.001
1000	11	0.3	-2	0.25	-0.017	0.009	0.006	-0.017	-0.001
1000	11	0.3	-2	0.5	-0.024	0	0.003	-0.024	0.002

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
1000	11	0.3	-2	0.8	-0.02	0.017	0.005	-0.024	-0.003
1000	11	0.3	-1.5	0	-0.03	0.007	-0.011	-0.025	0.003
1000	11	0.3	-1.5	0.25	-0.028	0.009	0.008	-0.019	0.004
1000	11	0.3	-1.5	0.5	-0.032	0.018	-0.001	-0.034	0.003
1000	11	0.3	-1.5	0.8	-0.017	0.021	0.014	-0.019	0.013
1000	11	0.3	-1	0	-0.041	0.012	-0.025	-0.041	-0.003
1000	11	0.3	-1	0.25	-0.039	0.011	-0.001	-0.032	-0.001
1000	11	0.3	-1	0.5	-0.048	0	-0.02	-0.06	-0.004
1000	11	0.3	-1	0.8	-0.025	-0.003	0	-0.042	0.004
1000	11	0.3	-0.5	0	-0.059	0.018	-0.04	-0.058	-0.007
1000	11	0.3	-0.5	0.25	-0.036	0.034	-0.006	-0.039	0.001
1000	11	0.3	-0.5	0.5	-0.058	0.018	-0.03	-0.075	0.004
1000	11	0.3	-0.5	0.8	-0.057	0.017	-0.02	-0.071	0
1000	11	0.3	0	0	-0.079	0.052	-0.048	-0.068	0.001
1000	11	0.3	0	0.25	-0.069	0.043	-0.02	-0.057	-0.002
1000	11	0.3	0	0.5	-0.071	0.075	-0.065	-0.114	0.003
1000	11	0.3	0	0.8	-0.065	0.048	-0.049	-0.106	-0.001
1000	11	0.5	-2	0	-0.046	0.006	-0.024	-0.037	-0.007
1000	11	0.5	-2	0.25	-0.032	0.015	-0.006	-0.029	-0.002
1000	11	0.5	-2	0.5	-0.052	0	-0.019	-0.045	-0.006
1000	11	0.5	-2	0.8	-0.032	0.013	-0.007	-0.034	-0.005
1000	11	0.5	-1.5	0	-0.053	0.014	-0.032	-0.045	-0.004
1000	11	0.5	-1.5	0.25	-0.045	-0.011	-0.006	-0.032	0.004
1000	11	0.5	-1.5	0.5	-0.05	0.001	-0.023	-0.055	-0.002
1000	11	0.5	-1.5	0.8	-0.047	0.014	-0.016	-0.05	-0.007
1000	11	0.5	-1	0	-0.068	-0.015	-0.043	-0.059	-0.003
1000	11	0.5	-1	0.25	-0.056	0.016	-0.015	-0.046	-0.001
1000	11	0.5	-1	0.5	-0.082	-0.007	-0.038	-0.079	-0.004
1000	11	0.5	-1	0.8	-0.051	-0.003	-0.014	-0.056	0.006
1000	11	0.5	-0.5	0	-0.107	-0.002	-0.059	-0.081	-0.003
1000	11	0.5	-0.5	0.25	-0.081	0.011	-0.032	-0.067	-0.007
1000	11	0.5	-0.5	0.5	-0.095	-0.004	-0.047	-0.095	0.006
1000	11	0.5	-0.5	0.8	-0.08	0.021	-0.042	-0.094	-0.005
1000	11	0.5	0	0	-0.136	-0.011	-0.082	-0.109	-0.006
1000	11	0.5	0	0.25	-0.108	0.014	-0.039	-0.079	-0.002
1000	11	0.5	0	0.5	-0.131	0.047	-0.091	-0.145	0.001
1000	11	0.5	0	0.8	-0.115	0.043	-0.074	-0.135	0.001
1000	11	0.8	-2	0	-0.059	-0.033	-0.036	-0.047	-0.003
1000	11	0.8	-2	0.25	-0.048	-0.016	-0.024	-0.046	-0.007
1000	11	0.8	-2	0.5	-0.071	-0.013	-0.028	-0.053	0.003

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
1000	11	0.8	-2	0.8	-0.037	-0.031	-0.015	-0.039	0.003
1000	11	0.8	-1.5	0	-0.083	-0.034	-0.06	-0.073	-0.01
1000	11	0.8	-1.5	0.25	-0.076	-0.002	-0.033	-0.059	-0.005
1000	11	0.8	-1.5	0.5	-0.086	-0.035	-0.044	-0.075	0.001
1000	11	0.8	-1.5	0.8	-0.064	-0.023	-0.027	-0.058	0.003
1000	11	0.8	-1	0	-0.118	-0.047	-0.073	-0.092	-0.003
1000	11	0.8	-1	0.25	-0.094	-0.034	-0.042	-0.073	-0.002
1000	11	0.8	-1	0.5	-0.125	-0.04	-0.064	-0.105	-0.003
1000	11	0.8	-1	0.8	-0.086	-0.034	-0.044	-0.084	0.003
1000	11	0.8	-0.5	0	-0.162	-0.007	-0.097	-0.123	0
1000	11	0.8	-0.5	0.25	-0.131	-0.028	-0.058	-0.095	-0.003
1000	11	0.8	-0.5	0.5	-0.172	-0.012	-0.084	-0.138	0.004
1000	11	0.8	-0.5	0.8	-0.132	-0.024	-0.065	-0.119	0.003
1000	11	0.8	0	0	-0.225	-0.038	-0.133	-0.168	0
1000	11	0.8	0	0.25	-0.166	-0.051	-0.066	-0.112	0.006
1000	11	0.8	0	0.5	-0.217	0.367	-0.123	-0.186	0.006
1000	11	0.8	0	0.8	-0.188	-0.01	-0.105	-0.173	0.003
1000	22	0	-2	0	0.001	0.004	0.006	0.002	0.001
1000	22	0	-2	0.25	-0.006	0.018	0.002	-0.007	-0.008
1000	22	0	-2	0.5	0.001	0.008	0.011	0	0
1000	22	0	-2	0.8	0.005	0.004	0.016	0.002	0.002
1000	22	0	-1.5	0	0	0.017	-0.001	-0.005	-0.006
1000	22	0	-1.5	0.25	-0.001	0.007	0.01	-0.001	-0.001
1000	22	0	-1.5	0.5	0.008	0.006	0.022	0.007	0.008
1000	22	0	-1.5	0.8	0.002	0.01	0.017	-0.002	0
1000	22	0	-1	0	-0.003	0.014	0	-0.005	-0.004
1000	22	0	-1	0.25	-0.004	0.02	0.009	-0.005	-0.005
1000	22	0	-1	0.5	0.007	0.011	0.024	0.005	0.008
1000	22	0	-1	0.8	-0.001	0.012	0.022	-0.005	-0.001
1000	22	0	-0.5	0	0.006	0.012	0.005	0	0.002
1000	22	0	-0.5	0.25	0.002	0.014	0.012	-0.004	-0.001
1000	22	0	-0.5	0.5	0.001	0.021	0.017	-0.009	-0.001
1000	22	0	-0.5	0.8	0.001	0.021	0.026	-0.01	0.001
1000	22	0	0	0	0.002	0.019	0.001	-0.005	-0.003
1000	22	0	0	0.25	0.017	0.03	0.021	0.003	0.008
1000	22	0	0	0.5	0.006	0.033	0.013	-0.018	0.001
1000	22	0	0	0.8	0.002	0.04	0.022	-0.023	0.003
1000	22	0.3	-2	0	-0.015	0.005	-0.013	-0.015	-0.005
1000	22	0.3	-2	0.25	-0.011	-0.003	-0.005	-0.011	-0.001
1000	22	0.3	-2	0.5	-0.009	0.018	-0.001	-0.008	0.005

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
1000	22	0.3	-2	0.8	-0.008	0.017	-0.002	-0.012	0
1000	22	0.3	-1.5	0	-0.022	0.003	-0.015	-0.016	0
1000	22	0.3	-1.5	0.25	-0.015	-0.009	-0.005	-0.012	0.002
1000	22	0.3	-1.5	0.5	-0.026	0.005	-0.014	-0.025	-0.006
1000	22	0.3	-1.5	0.8	-0.017	0.005	-0.008	-0.023	-0.005
1000	22	0.3	-1	0	-0.034	0.005	-0.027	-0.028	-0.003
1000	22	0.3	-1	0.25	-0.021	0.024	-0.005	-0.014	0.006
1000	22	0.3	-1	0.5	-0.034	0.019	-0.013	-0.029	-0.001
1000	22	0.3	-1	0.8	-0.021	0.014	-0.001	-0.023	0.004
1000	22	0.3	-0.5	0	-0.048	0	-0.039	-0.042	-0.007
1000	22	0.3	-0.5	0.25	-0.033	0.009	-0.019	-0.031	-0.003
1000	22	0.3	-0.5	0.5	-0.044	0.019	-0.014	-0.036	0.005
1000	22	0.3	-0.5	0.8	-0.032	0.011	-0.011	-0.042	-0.001
1000	22	0.3	0	0	-0.047	0.006	-0.032	-0.036	0.008
1000	22	0.3	0	0.25	-0.057	0.012	-0.024	-0.04	-0.001
1000	22	0.3	0	0.5	-0.057	0.007	-0.028	-0.059	-0.002
1000	22	0.3	0	0.8	-0.053	0.005	-0.02	-0.063	0
1000	22	0.5	-2	0	-0.02	0.001	-0.019	-0.019	0
1000	22	0.5	-2	0.25	-0.016	-0.002	-0.008	-0.013	0.004
1000	22	0.5	-2	0.5	-0.026	0	-0.02	-0.027	-0.005
1000	22	0.5	-2	0.8	-0.018	0.005	-0.009	-0.018	0.002
1000	22	0.5	-1.5	0	-0.034	0.001	-0.027	-0.026	0.001
1000	22	0.5	-1.5	0.25	-0.037	0.009	-0.019	-0.025	-0.001
1000	22	0.5	-1.5	0.5	-0.042	0.001	-0.027	-0.036	-0.004
1000	22	0.5	-1.5	0.8	-0.038	0.003	-0.024	-0.037	-0.007
1000	22	0.5	-1	0	-0.049	-0.013	-0.041	-0.039	0.002
1000	22	0.5	-1	0.25	-0.029	-0.012	-0.02	-0.025	0.01
1000	22	0.5	-1	0.5	-0.055	-0.001	-0.037	-0.05	-0.004
1000	22	0.5	-1	0.8	-0.043	0.002	-0.02	-0.038	0.003
1000	22	0.5	-0.5	0	-0.074	-0.004	-0.06	-0.06	-0.004
1000	22	0.5	-0.5	0.25	-0.064	0.004	-0.042	-0.051	-0.004
1000	22	0.5	-0.5	0.5	-0.072	-0.003	-0.04	-0.06	0.004
1000	22	0.5	-0.5	0.8	-0.064	-0.006	-0.031	-0.059	0.002
1000	22	0.5	0	0	-0.109	0.006	-0.079	-0.083	-0.005
1000	22	0.5	0	0.25	-0.089	0.012	-0.047	-0.06	0.001
1000	22	0.5	0	0.5	-0.104	0.01	-0.063	-0.092	-0.007
1000	22	0.5	0	0.8	-0.092	0.002	-0.048	-0.09	-0.001
1000	22	0.8	-2	0	-0.03	-0.019	-0.027	-0.024	0.006
1000	22	0.8	-2	0.25	-0.035	-0.008	-0.027	-0.029	-0.003
1000	22	0.8	-2	0.5	-0.034	-0.011	-0.025	-0.028	0.006

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
1000	22	0.8	-2	0.8	-0.029	-0.005	-0.022	-0.027	0.003
1000	22	0.8	-1.5	0	-0.055	-0.021	-0.054	-0.049	-0.002
1000	22	0.8	-1.5	0.25	-0.041	-0.015	-0.035	-0.036	0.003
1000	22	0.8	-1.5	0.5	-0.056	-0.025	-0.043	-0.047	0.002
1000	22	0.8	-1.5	0.8	-0.051	-0.001	-0.036	-0.045	-0.001
1000	22	0.8	-1	0	-0.071	-0.036	-0.069	-0.062	0.007
1000	22	0.8	-1	0.25	-0.071	-0.015	-0.051	-0.053	0.003
1000	22	0.8	-1	0.5	-0.086	-0.021	-0.061	-0.07	0
1000	22	0.8	-1	0.8	-0.075	-0.029	-0.053	-0.066	0.001
1000	22	0.8	-0.5	0	-0.123	-0.043	-0.104	-0.1	-0.006
1000	22	0.8	-0.5	0.25	-0.095	-0.037	-0.071	-0.074	0
1000	22	0.8	-0.5	0.5	-0.119	-0.023	-0.08	-0.095	0.004
1000	22	0.8	-0.5	0.8	-0.105	-0.019	-0.068	-0.091	0.005
1000	22	0.8	0	0	-0.161	-0.051	-0.126	-0.127	0
1000	22	0.8	0	0.25	-0.127	-0.04	-0.086	-0.094	0.004
1000	22	0.8	0	0.5	-0.159	-0.049	-0.103	-0.129	0.005
1000	22	0.8	0	0.8	-0.162	-0.025	-0.1	-0.136	-0.006
1000	33	0	-2	0	0.005	0.014	0.003	0.003	0.002
1000	33	0	-2	0.25	0.003	0.007	0.007	0.005	0.005
1000	33	0	-2	0.5	-0.002	0.009	0.004	-0.002	-0.003
1000	33	0	-2	0.8	0.009	0.016	0.016	0.008	0.009
1000	33	0	-1.5	0	0	0.001	0.002	0.002	0.001
1000	33	0	-1.5	0.25	0.003	0.014	0.002	-0.002	-0.003
1000	33	0	-1.5	0.5	0	0.015	0.005	-0.005	-0.005
1000	33	0	-1.5	0.8	-0.003	0.01	0.007	-0.005	-0.005
1000	33	0	-1	0	-0.002	-0.001	-0.002	-0.003	-0.003
1000	33	0	-1	0.25	-0.003	0.007	0	-0.007	-0.006
1000	33	0	-1	0.5	0	0.015	0.013	-0.001	0
1000	33	0	-1	0.8	-0.01	0.017	0.01	-0.009	-0.007
1000	33	0	-0.5	0	-0.005	0.015	0.004	0.002	0.001
1000	33	0	-0.5	0.25	-0.004	0.007	0.008	-0.001	-0.001
1000	33	0	-0.5	0.5	0.004	0.008	0.018	-0.002	0.001
1000	33	0	-0.5	0.8	-0.002	0.021	0.019	-0.008	-0.003
1000	33	0	0	0	0	0.01	-0.004	-0.005	-0.003
1000	33	0	0	0.25	-0.007	0.032	0.004	-0.008	-0.005
1000	33	0	0	0.5	-0.008	0.026	0.01	-0.016	-0.01
1000	33	0	0	0.8	-0.004	0.01	0.021	-0.016	-0.006
1000	33	0.3	-2	0	-0.002	0.001	-0.004	-0.002	0.006
1000	33	0.3	-2	0.25	-0.007	0.003	-0.008	-0.009	0.001
1000	33	0.3	-2	0.5	-0.006	0.012	-0.001	-0.005	0.006

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
1000	33	0.3	-2	0.8	0.003	0.008	0.004	-0.001	0.008
1000	33	0.3	-1.5	0	-0.014	0.01	-0.015	-0.013	0.001
1000	33	0.3	-1.5	0.25	-0.009	0.005	-0.007	-0.008	0.005
1000	33	0.3	-1.5	0.5	-0.019	0.007	-0.014	-0.019	-0.004
1000	33	0.3	-1.5	0.8	-0.015	0.01	-0.006	-0.015	-0.001
1000	33	0.3	-1	0	-0.023	0.005	-0.024	-0.021	-0.001
1000	33	0.3	-1	0.25	-0.031	0.01	-0.026	-0.029	-0.007
1000	33	0.3	-1	0.5	-0.022	0.009	-0.013	-0.022	0
1000	33	0.3	-1	0.8	-0.015	0.008	-0.001	-0.014	0.007
1000	33	0.3	-0.5	0	-0.04	0.013	-0.036	-0.032	-0.004
1000	33	0.3	-0.5	0.25	-0.038	0.014	-0.029	-0.034	-0.004
1000	33	0.3	-0.5	0.5	-0.035	0.025	-0.018	-0.032	-0.003
1000	33	0.3	-0.5	0.8	-0.026	0.016	-0.005	-0.027	0.003
1000	33	0.3	0	0	-0.057	-0.003	-0.045	-0.042	-0.005
1000	33	0.3	0	0.25	-0.062	0.022	-0.043	-0.052	-0.012
1000	33	0.3	0	0.5	-0.049	0.026	-0.026	-0.046	-0.002
1000	33	0.3	0	0.8	-0.055	0.012	-0.017	-0.05	-0.007
1000	33	0.5	-2	0	-0.014	-0.003	-0.018	-0.016	-0.001
1000	33	0.5	-2	0.25	-0.019	0.005	-0.016	-0.016	0
1000	33	0.5	-2	0.5	-0.021	0.006	-0.014	-0.017	0
1000	33	0.5	-2	0.8	-0.014	0.005	-0.011	-0.015	-0.001
1000	33	0.5	-1.5	0	-0.023	0.011	-0.028	-0.023	-0.001
1000	33	0.5	-1.5	0.25	-0.025	0.002	-0.021	-0.019	0.005
1000	33	0.5	-1.5	0.5	-0.032	-0.002	-0.025	-0.028	-0.003
1000	33	0.5	-1.5	0.8	-0.022	0.001	-0.011	-0.017	0.005
1000	33	0.5	-1	0	-0.045	-0.001	-0.048	-0.042	-0.009
1000	33	0.5	-1	0.25	-0.038	-0.003	-0.037	-0.037	-0.002
1000	33	0.5	-1	0.5	-0.043	0.007	-0.035	-0.04	-0.004
1000	33	0.5	-1	0.8	-0.032	0.002	-0.019	-0.03	0.003
1000	33	0.5	-0.5	0	-0.066	-0.006	-0.063	-0.056	-0.008
1000	33	0.5	-0.5	0.25	-0.055	-0.001	-0.046	-0.047	0
1000	33	0.5	-0.5	0.5	-0.061	0.003	-0.045	-0.055	-0.004
1000	33	0.5	-0.5	0.8	-0.05	0.004	-0.028	-0.046	0.002
1000	33	0.5	0	0	-0.075	-0.008	-0.066	-0.058	0.005
1000	33	0.5	0	0.25	-0.075	-0.002	-0.058	-0.062	0.002
1000	33	0.5	0	0.5	-0.075	0.003	-0.044	-0.06	0.009
1000	33	0.5	0	0.8	-0.08	0.012	-0.044	-0.072	-0.002
1000	33	0.8	-2	0	-0.029	-0.021	-0.031	-0.027	-0.003
1000	33	0.8	-2	0.25	-0.026	-0.014	-0.029	-0.026	0.001
1000	33	0.8	-2	0.5	-0.033	-0.007	-0.028	-0.029	-0.002

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
1000	33	0.8	-2	0.8	-0.021	-0.005	-0.019	-0.021	0.002
1000	33	0.8	-1.5	0	-0.045	-0.026	-0.05	-0.043	-0.006
1000	33	0.8	-1.5	0.25	-0.044	-0.011	-0.04	-0.036	0.002
1000	33	0.8	-1.5	0.5	-0.051	-0.016	-0.048	-0.048	-0.008
1000	33	0.8	-1.5	0.8	-0.042	-0.007	-0.035	-0.038	-0.002
1000	33	0.8	-1	0	-0.059	-0.031	-0.066	-0.056	-0.001
1000	33	0.8	-1	0.25	-0.067	-0.031	-0.06	-0.056	0
1000	33	0.8	-1	0.5	-0.062	-0.025	-0.052	-0.053	0.003
1000	33	0.8	-1	0.8	-0.057	-0.021	-0.047	-0.053	0
1000	33	0.8	-0.5	0	-0.092	-0.039	-0.092	-0.079	-0.001
1000	33	0.8	-0.5	0.25	-0.088	-0.022	-0.077	-0.073	0.006
1000	33	0.8	-0.5	0.5	-0.091	-0.038	-0.07	-0.073	0.007
1000	33	0.8	-0.5	0.8	-0.087	-0.022	-0.066	-0.078	-0.003
1000	33	0.8	0	0	-0.134	-0.043	-0.12	-0.108	-0.001
1000	33	0.8	0	0.25	-0.126	-0.037	-0.105	-0.104	0.003
1000	33	0.8	0	0.5	-0.14	-0.028	-0.102	-0.113	-0.003
1000	33	0.8	0	0.8	-0.122	-0.031	-0.082	-0.103	0.004
2000	11	0	-2	0	0.008	0.003	0.017	0.003	0.006
2000	11	0	-2	0.25	0	0.014	0.018	-0.006	-0.002
2000	11	0	-2	0.5	0	0.003	0.019	-0.008	0.001
2000	11	0	-2	0.8	-0.001	0.004	0.021	-0.009	-0.001
2000	11	0	-1.5	0	-0.001	0.014	0.01	-0.004	-0.001
2000	11	0	-1.5	0.25	-0.003	0.01	0.022	-0.005	0
2000	11	0	-1.5	0.5	-0.001	0.02	0.015	-0.017	-0.001
2000	11	0	-1.5	0.8	-0.007	0.009	0.018	-0.018	-0.005
2000	11	0	-1	0	-0.001	0.019	0.005	-0.009	-0.002
2000	11	0	-1	0.25	0	0.003	0.019	-0.011	-0.002
2000	11	0	-1	0.5	0	0.014	0.004	-0.033	-0.004
2000	11	0	-1	0.8	0.005	0.016	0.025	-0.017	0.006
2000	11	0	-0.5	0	0	0.011	0.002	-0.011	0.001
2000	11	0	-0.5	0.25	-0.004	0.007	0.016	-0.016	-0.004
2000	11	0	-0.5	0.5	-0.004	0.006	-0.002	-0.043	0.004
2000	11	0	-0.5	0.8	0.002	0.022	0.004	-0.044	-0.002
2000	11	0	0	0	0.003	0.016	-0.005	-0.017	-0.001
2000	11	0	0	0.25	0.001	0.015	0.012	-0.019	0
2000	11	0	0	0.5	0.007	0.029	-0.028	-0.069	0.007
2000	11	0	0	0.8	0.004	0.019	-0.015	-0.065	0.006
2000	11	0.3	-2	0	-0.02	0.004	-0.005	-0.018	0
2000	11	0.3	-2	0.25	-0.013	0.001	0.007	-0.015	0.001
2000	11	0.3	-2	0.5	-0.017	-0.005	0.006	-0.02	0.005

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
2000	11	0.3	-2	0.8	-0.018	0.002	0.004	-0.023	-0.004
2000	11	0.3	-1.5	0	-0.024	-0.002	-0.013	-0.026	0
2000	11	0.3	-1.5	0.25	-0.011	0.007	0.009	-0.016	0.007
2000	11	0.3	-1.5	0.5	-0.03	-0.008	-0.001	-0.033	0.004
2000	11	0.3	-1.5	0.8	-0.025	0.001	0.005	-0.029	0.001
2000	11	0.3	-1	0	-0.04	0.002	-0.02	-0.035	0.001
2000	11	0.3	-1	0.25	-0.029	0.002	-0.001	-0.03	0.001
2000	11	0.3	-1	0.5	-0.044	0.009	-0.016	-0.055	-0.001
2000	11	0.3	-1	0.8	-0.037	0.007	-0.006	-0.049	-0.001
2000	11	0.3	-0.5	0	-0.061	0.007	-0.033	-0.051	-0.001
2000	11	0.3	-0.5	0.25	-0.048	0.006	-0.006	-0.04	0.001
2000	11	0.3	-0.5	0.5	-0.061	0.011	-0.036	-0.082	-0.001
2000	11	0.3	-0.5	0.8	-0.042	0.008	-0.018	-0.068	0.004
2000	11	0.3	0	0	-0.08	0.015	-0.047	-0.067	0.002
2000	11	0.3	0	0.25	-0.065	0.014	-0.017	-0.054	-0.002
2000	11	0.3	0	0.5	-0.08	0.021	-0.064	-0.113	0.004
2000	11	0.3	0	0.8	-0.07	0.015	-0.052	-0.109	0
2000	11	0.5	-2	0	-0.033	-0.011	-0.017	-0.029	0
2000	11	0.5	-2	0.25	-0.032	-0.008	-0.003	-0.026	0
2000	11	0.5	-2	0.5	-0.041	-0.001	-0.012	-0.038	-0.001
2000	11	0.5	-2	0.8	-0.036	-0.008	-0.005	-0.032	-0.004
2000	11	0.5	-1.5	0	-0.052	-0.022	-0.031	-0.045	-0.005
2000	11	0.5	-1.5	0.25	-0.042	-0.002	-0.008	-0.034	0.001
2000	11	0.5	-1.5	0.5	-0.063	0	-0.024	-0.057	-0.005
2000	11	0.5	-1.5	0.8	-0.04	-0.006	-0.01	-0.043	-0.001
2000	11	0.5	-1	0	-0.075	-0.006	-0.043	-0.06	-0.002
2000	11	0.5	-1	0.25	-0.06	-0.011	-0.017	-0.048	-0.002
2000	11	0.5	-1	0.5	-0.072	-0.011	-0.033	-0.073	0
2000	11	0.5	-1	0.8	-0.063	0	-0.023	-0.066	-0.004
2000	11	0.5	-0.5	0	-0.096	-0.011	-0.056	-0.077	0.002
2000	11	0.5	-0.5	0.25	-0.084	-0.01	-0.029	-0.065	-0.004
2000	11	0.5	-0.5	0.5	-0.116	0.006	-0.064	-0.113	-0.008
2000	11	0.5	-0.5	0.8	-0.087	-0.002	-0.038	-0.091	0.002
2000	11	0.5	0	0	-0.14	-0.024	-0.086	-0.112	-0.005
2000	11	0.5	0	0.25	-0.114	-0.002	-0.044	-0.084	-0.006
2000	11	0.5	0	0.5	-0.137	0.007	-0.094	-0.149	-0.004
2000	11	0.5	0	0.8	-0.114	-0.01	-0.069	-0.13	0.003
2000	11	0.8	-2	0	-0.06	-0.027	-0.037	-0.048	-0.002
2000	11	0.8	-2	0.25	-0.05	-0.026	-0.021	-0.042	-0.003
2000	11	0.8	-2	0.5	-0.067	-0.033	-0.033	-0.057	-0.002

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
2000	11	0.8	-2	0.8	-0.04	-0.019	-0.017	-0.041	0
2000	11	0.8	-1.5	0	-0.078	-0.038	-0.051	-0.064	-0.001
2000	11	0.8	-1.5	0.25	-0.06	-0.024	-0.027	-0.051	0.004
2000	11	0.8	-1.5	0.5	-0.091	-0.038	-0.046	-0.077	-0.002
2000	11	0.8	-1.5	0.8	-0.064	-0.031	-0.027	-0.058	0.002
2000	11	0.8	-1	0	-0.116	-0.063	-0.07	-0.088	0.001
2000	11	0.8	-1	0.25	-0.097	-0.036	-0.044	-0.075	-0.003
2000	11	0.8	-1	0.5	-0.121	-0.056	-0.057	-0.098	0.004
2000	11	0.8	-1	0.8	-0.098	-0.036	-0.045	-0.087	0
2000	11	0.8	-0.5	0	-0.165	-0.054	-0.096	-0.122	0.001
2000	11	0.8	-0.5	0.25	-0.134	-0.048	-0.057	-0.095	-0.002
2000	11	0.8	-0.5	0.5	-0.16	-0.048	-0.084	-0.137	0.001
2000	11	0.8	-0.5	0.8	-0.136	-0.031	-0.069	-0.124	0.001
2000	11	0.8	0	0	-0.219	-0.044	-0.126	-0.161	0.004
2000	11	0.8	0	0.25	-0.177	-0.039	-0.071	-0.116	-0.001
2000	11	0.8	0	0.5	-0.215	0.004	-0.123	-0.187	0.004
2000	11	0.8	0	0.8	-0.194	-0.025	-0.11	-0.179	-0.004
2000	22	0	-2	0	-0.002	0.007	0.002	-0.002	-0.003
2000	22	0	-2	0.25	0.004	0.013	0.014	0.005	0.004
2000	22	0	-2	0.5	-0.001	0.006	0.01	-0.002	-0.001
2000	22	0	-2	0.8	0.001	0.003	0.016	0.002	0.003
2000	22	0	-1.5	0	0	0.008	0.003	-0.001	-0.001
2000	22	0	-1.5	0.25	-0.004	0.002	0.01	-0.001	-0.001
2000	22	0	-1.5	0.5	0.004	0.002	0.016	0.001	0.003
2000	22	0	-1.5	0.8	-0.004	0.002	0.011	-0.009	-0.007
2000	22	0	-1	0	-0.004	0.011	0.002	-0.003	-0.003
2000	22	0	-1	0.25	-0.003	0.015	0.012	-0.002	-0.001
2000	22	0	-1	0.5	0	0.006	0.015	-0.004	-0.001
2000	22	0	-1	0.8	0	0.004	0.021	-0.006	-0.002
2000	22	0	-0.5	0	-0.002	0.011	0.002	-0.004	-0.003
2000	22	0	-0.5	0.25	0	0.013	0.013	-0.003	-0.002
2000	22	0	-0.5	0.5	-0.007	0.008	0.012	-0.014	-0.006
2000	22	0	-0.5	0.8	-0.001	0.008	0.023	-0.013	-0.002
2000	22	0	0	0	0	0.012	0.001	-0.005	-0.001
2000	22	0	0	0.25	0.008	0.012	0.021	0.003	0.007
2000	22	0	0	0.5	-0.001	0.008	0.009	-0.022	-0.003
2000	22	0	0	0.8	0.002	0.007	0.017	-0.028	-0.002
2000	22	0.3	-2	0	-0.013	-0.002	-0.011	-0.013	-0.001
2000	22	0.3	-2	0.25	-0.007	0.003	-0.002	-0.008	0.002
2000	22	0.3	-2	0.5	-0.013	-0.002	-0.003	-0.011	0.001

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
2000	22	0.3	-2	0.8	-0.014	0.003	-0.003	-0.014	-0.002
2000	22	0.3	-1.5	0	-0.019	-0.004	-0.016	-0.017	0
2000	22	0.3	-1.5	0.25	-0.025	-0.002	-0.009	-0.017	-0.003
2000	22	0.3	-1.5	0.5	-0.022	0.002	-0.007	-0.019	0
2000	22	0.3	-1.5	0.8	-0.013	0.001	0.001	-0.014	0.004
2000	22	0.3	-1	0	-0.032	0.003	-0.024	-0.025	0
2000	22	0.3	-1	0.25	-0.033	0.003	-0.019	-0.029	-0.008
2000	22	0.3	-1	0.5	-0.034	-0.001	-0.017	-0.032	-0.003
2000	22	0.3	-1	0.8	-0.033	0.007	-0.011	-0.033	-0.005
2000	22	0.3	-0.5	0	-0.048	-0.003	-0.035	-0.038	-0.003
2000	22	0.3	-0.5	0.25	-0.033	0.004	-0.015	-0.027	0.002
2000	22	0.3	-0.5	0.5	-0.04	-0.006	-0.016	-0.038	0.002
2000	22	0.3	-0.5	0.8	-0.039	0.003	-0.01	-0.042	0.001
2000	22	0.3	0	0	-0.06	0.005	-0.041	-0.046	-0.001
2000	22	0.3	0	0.25	-0.044	0	-0.018	-0.033	0.005
2000	22	0.3	0	0.5	-0.061	0.003	-0.03	-0.06	-0.001
2000	22	0.3	0	0.8	-0.051	0.008	-0.017	-0.059	0.002
2000	22	0.5	-2	0	-0.022	0.004	-0.02	-0.02	-0.001
2000	22	0.5	-2	0.25	-0.018	-0.007	-0.008	-0.012	0.005
2000	22	0.5	-2	0.5	-0.02	-0.004	-0.014	-0.02	0.001
2000	22	0.5	-2	0.8	-0.022	-0.004	-0.012	-0.021	-0.002
2000	22	0.5	-1.5	0	-0.033	-0.005	-0.031	-0.029	-0.001
2000	22	0.5	-1.5	0.25	-0.03	-0.012	-0.019	-0.024	0
2000	22	0.5	-1.5	0.5	-0.037	-0.006	-0.022	-0.031	0.001
2000	22	0.5	-1.5	0.8	-0.033	-0.008	-0.015	-0.028	0.001
2000	22	0.5	-1	0	-0.046	-0.019	-0.042	-0.04	0.001
2000	22	0.5	-1	0.25	-0.049	-0.004	-0.032	-0.038	-0.004
2000	22	0.5	-1	0.5	-0.045	-0.007	-0.026	-0.039	0.005
2000	22	0.5	-1	0.8	-0.045	-0.008	-0.02	-0.039	0.003
2000	22	0.5	-0.5	0	-0.068	-0.017	-0.054	-0.054	0.003
2000	22	0.5	-0.5	0.25	-0.061	-0.007	-0.036	-0.045	0.001
2000	22	0.5	-0.5	0.5	-0.071	-0.001	-0.041	-0.061	0.001
2000	22	0.5	-0.5	0.8	-0.063	-0.005	-0.033	-0.061	0.001
2000	22	0.5	0	0	-0.096	-0.015	-0.07	-0.074	0.002
2000	22	0.5	0	0.25	-0.082	-0.006	-0.043	-0.056	0.004
2000	22	0.5	0	0.5	-0.108	-0.01	-0.059	-0.089	-0.002
2000	22	0.5	0	0.8	-0.09	0.002	-0.043	-0.083	0.003
2000	22	0.8	-2	0	-0.039	-0.013	-0.038	-0.036	-0.005
2000	22	0.8	-2	0.25	-0.029	-0.019	-0.023	-0.024	0.003
2000	22	0.8	-2	0.5	-0.037	-0.021	-0.03	-0.033	0.001

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
2000	22	0.8	-2	0.8	-0.036	-0.019	-0.026	-0.032	-0.002
2000	22	0.8	-1.5	0	-0.048	-0.027	-0.05	-0.045	0.001
2000	22	0.8	-1.5	0.25	-0.052	-0.025	-0.041	-0.043	-0.004
2000	22	0.8	-1.5	0.5	-0.059	-0.03	-0.045	-0.05	0
2000	22	0.8	-1.5	0.8	-0.051	-0.027	-0.038	-0.046	-0.001
2000	22	0.8	-1	0	-0.083	-0.041	-0.069	-0.064	0
2000	22	0.8	-1	0.25	-0.066	-0.028	-0.048	-0.049	0.005
2000	22	0.8	-1	0.5	-0.081	-0.025	-0.066	-0.074	-0.002
2000	22	0.8	-1	0.8	-0.076	-0.028	-0.052	-0.065	0
2000	22	0.8	-0.5	0	-0.116	-0.053	-0.093	-0.089	0.003
2000	22	0.8	-0.5	0.25	-0.098	-0.036	-0.07	-0.073	0.001
2000	22	0.8	-0.5	0.5	-0.118	-0.042	-0.081	-0.096	0.002
2000	22	0.8	-0.5	0.8	-0.103	-0.048	-0.066	-0.088	0.005
2000	22	0.8	0	0	-0.163	-0.06	-0.121	-0.122	0.003
2000	22	0.8	0	0.25	-0.135	-0.036	-0.086	-0.094	0.005
2000	22	0.8	0	0.5	-0.168	-0.04	-0.107	-0.133	0
2000	22	0.8	0	0.8	-0.156	-0.039	-0.095	-0.131	-0.002
2000	33	0	-2	0	0	0.006	0.001	0	0
2000	33	0	-2	0.25	0	0.009	0.002	0	-0.001
2000	33	0	-2	0.5	-0.002	0.005	0.007	0	0
2000	33	0	-2	0.8	0.002	-0.003	0.011	0.003	0.003
2000	33	0	-1.5	0	0.004	0	0.003	0.002	0.001
2000	33	0	-1.5	0.25	-0.001	0.01	-0.001	-0.006	-0.006
2000	33	0	-1.5	0.5	0.004	0.008	0.014	0.005	0.004
2000	33	0	-1.5	0.8	0.002	0.004	0.012	0	0.001
2000	33	0	-1	0	-0.003	0.006	0.001	0	-0.001
2000	33	0	-1	0.25	0.003	0.007	0.005	-0.001	-0.001
2000	33	0	-1	0.5	0.003	0.011	0.014	0.001	0.001
2000	33	0	-1	0.8	0.004	0.004	0.021	0.002	0.004
2000	33	0	-0.5	0	0.002	0.011	0.006	0.005	0.005
2000	33	0	-0.5	0.25	-0.004	0.001	0.005	-0.004	-0.004
2000	33	0	-0.5	0.5	0.007	0.011	0.021	0.002	0.005
2000	33	0	-0.5	0.8	-0.002	0.012	0.019	-0.009	-0.005
2000	33	0	0	0	0.002	0.015	0.003	0.002	0.001
2000	33	0	0	0.25	0.003	0.012	0.007	-0.005	-0.001
2000	33	0	0	0.5	0.004	0.004	0.016	-0.009	-0.001
2000	33	0	0	0.8	-0.005	0.01	0.023	-0.015	-0.003
2000	33	0.3	-2	0	-0.009	0.008	-0.009	-0.008	0.001
2000	33	0.3	-2	0.25	-0.007	0.004	-0.008	-0.009	0.001
2000	33	0.3	-2	0.5	-0.007	0.004	-0.004	-0.008	0.002

Design					Mean Bias				
N	I	$Cor(\xi,\theta)$	$Miss.$	$r(\beta,\gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
2000	33	0.3	-2	0.8	-0.007	0.001	-0.002	-0.007	0.002
2000	33	0.3	-1.5	0	-0.013	0.007	-0.017	-0.015	-0.002
2000	33	0.3	-1.5	0.25	-0.015	0.008	-0.01	-0.011	0.003
2000	33	0.3	-1.5	0.5	-0.024	0.003	-0.016	-0.022	-0.007
2000	33	0.3	-1.5	0.8	-0.012	-0.001	-0.004	-0.012	0.002
2000	33	0.3	-1	0	-0.028	-0.007	-0.026	-0.023	-0.002
2000	33	0.3	-1	0.25	-0.022	-0.004	-0.017	-0.019	0.001
2000	33	0.3	-1	0.5	-0.025	0.003	-0.014	-0.023	-0.001
2000	33	0.3	-1	0.8	-0.021	0.007	-0.006	-0.02	0
2000	33	0.3	-0.5	0	-0.032	-0.002	-0.031	-0.027	0.002
2000	33	0.3	-0.5	0.25	-0.034	0.004	-0.024	-0.028	0
2000	33	0.3	-0.5	0.5	-0.038	-0.003	-0.02	-0.034	-0.002
2000	33	0.3	-0.5	0.8	-0.027	0.011	-0.007	-0.028	0.003
2000	33	0.3	0	0	-0.049	0.004	-0.042	-0.038	0
2000	33	0.3	0	0.25	-0.049	0	-0.035	-0.043	-0.001
2000	33	0.3	0	0.5	-0.052	0.012	-0.03	-0.05	-0.004
2000	33	0.3	0	0.8	-0.05	0.017	-0.016	-0.048	-0.003
2000	33	0.5	-2	0	-0.018	-0.005	-0.021	-0.018	-0.003
2000	33	0.5	-2	0.25	-0.022	-0.009	-0.02	-0.02	-0.003
2000	33	0.5	-2	0.5	-0.016	0.003	-0.014	-0.016	0.001
2000	33	0.5	-2	0.8	-0.012	-0.007	-0.004	-0.008	0.006
2000	33	0.5	-1.5	0	-0.027	-0.002	-0.028	-0.024	-0.001
2000	33	0.5	-1.5	0.25	-0.034	-0.012	-0.029	-0.028	-0.004
2000	33	0.5	-1.5	0.5	-0.026	0.003	-0.019	-0.022	0.003
2000	33	0.5	-1.5	0.8	-0.021	-0.006	-0.012	-0.018	0.004
2000	33	0.5	-1	0	-0.039	-0.014	-0.042	-0.036	-0.002
2000	33	0.5	-1	0.25	-0.046	-0.003	-0.039	-0.039	-0.005
2000	33	0.5	-1	0.5	-0.036	-0.005	-0.026	-0.031	0.004
2000	33	0.5	-1	0.8	-0.036	-0.003	-0.022	-0.032	0
2000	33	0.5	-0.5	0	-0.05	-0.014	-0.052	-0.044	0.004
2000	33	0.5	-0.5	0.25	-0.06	-0.002	-0.048	-0.049	-0.001
2000	33	0.5	-0.5	0.5	-0.061	-0.003	-0.04	-0.051	0
2000	33	0.5	-0.5	0.8	-0.048	-0.007	-0.032	-0.049	-0.002
2000	33	0.5	0	0	-0.081	-0.011	-0.074	-0.067	-0.002
2000	33	0.5	0	0.25	-0.078	-0.014	-0.062	-0.067	-0.002
2000	33	0.5	0	0.5	-0.081	-0.007	-0.053	-0.069	-0.001
2000	33	0.5	0	0.8	-0.074	-0.005	-0.041	-0.069	0
2000	33	0.8	-2	0	-0.027	-0.01	-0.03	-0.026	-0.002
2000	33	0.8	-2	0.25	-0.031	-0.012	-0.034	-0.032	-0.006
2000	33	0.8	-2	0.5	-0.029	-0.01	-0.025	-0.025	0.002

Design					Mean Bias				
N	I	$Cor(\xi, \theta)$	$Miss.$	$r(\beta, \gamma)$	$\hat{\beta}_i$	$\hat{\alpha}_i$	$\hat{\xi}_{ML}$	$\hat{\xi}_{WML}$	$\hat{\xi}_{EAP}$
2000	33	0.8	-2	0.8	-0.023	-0.023	-0.018	-0.02	0.004
2000	33	0.8	-1.5	0	-0.04	-0.03	-0.046	-0.039	-0.002
2000	33	0.8	-1.5	0.25	-0.048	-0.015	-0.049	-0.046	-0.008
2000	33	0.8	-1.5	0.5	-0.043	-0.014	-0.04	-0.04	0
2000	33	0.8	-1.5	0.8	-0.036	-0.017	-0.029	-0.032	0.003
2000	33	0.8	-1	0	-0.061	-0.026	-0.065	-0.055	0
2000	33	0.8	-1	0.25	-0.065	-0.028	-0.06	-0.056	0
2000	33	0.8	-1	0.5	-0.068	-0.029	-0.06	-0.061	-0.004
2000	33	0.8	-1	0.8	-0.059	-0.029	-0.047	-0.052	0
2000	33	0.8	-0.5	0	-0.096	-0.042	-0.095	-0.083	-0.002
2000	33	0.8	-0.5	0.25	-0.093	-0.039	-0.083	-0.079	0
2000	33	0.8	-0.5	0.5	-0.089	-0.036	-0.071	-0.075	0.005
2000	33	0.8	-0.5	0.8	-0.08	-0.03	-0.062	-0.073	0.002
2000	33	0.8	0	0	-0.132	-0.057	-0.121	-0.109	0
2000	33	0.8	0	0.25	-0.138	-0.049	-0.109	-0.109	-0.003
2000	33	0.8	0	0.5	-0.131	-0.037	-0.095	-0.105	0.005
2000	33	0.8	0	0.8	-0.124	-0.048	-0.085	-0.106	-0.001

Appendix B

In this Appendix, the input files of ConQuest (Wu et al., 1998) and M*plus* (Muthén & Muthén, 1998 - 2010) are listed that were used for the analyses of Data Example A. Comments in the ConQuest syntax starts with „/*“ and ends „*/“ . In M*plus* comments starts with „!“.

One-parameter and Rasch-equivalent MIRT models The B-MIRT Rasch model can be applied using software for multidimensional Rasch-models, such as ConQuest. Listing A.1 shows the ConQuest input file of the B-MIRT Rasch model used for Data Example A. The ConQuest input file of W_{Dif}-MIRT Rasch model is given in Listing A.2. Note that this model can only be specified in ConQuest if all parameters γ^*_{im} of $\boldsymbol{\gamma}^*_{\xi}$ are zero or one. However, in general, $\gamma^*_{im} = \sum_{l=1}^{P} \gamma_{il}$ in the W_{Dif}Rasch model. Hence, γ^*_{im} is only zero or one if all variables D_i indicate only one latent response propensity θ_l. If at least a single indicator variable D_i indicates more than one latent dimension θ_l, then $\gamma^*_{im} = j$, with $j \in \{2, \ldots, P\}$. In this case, software for two-parameter MIRT models needs to be used.

Listing A.1: ConQuest input file for the B-MIRT Rasch Model (Data Example A).

```
datafile DataExampleA.dat;
format id 1-5 responses 6-65;
codes 0,1;
set update=yes,warnings=no;
score (0,1) (0,1) (   ) ! items(1-30);  /* Items Yi*/
score (0,1) (   ) (0,1) ! items(31-60); /* Response Indicators Di*/
model items;
/* Model parameter estimates  */
export parameters       >> between.prm; /* Item difficulties */
export covariance       >> between.cov; /* latent (co-)variances */
/* Starting estimation & call for item fit (In-/Outfit) */
estimate, fit=yes;
/* Person parameter estimates */
show cases ! estimates = latent >> between.lat; /* EAPs & PVs   */
show cases ! estimates = mle    >> between.mle; /* ML estimates */
show cases ! estimates = wle    >> between.wle; /* WML estimates*/
/* Output */
show       ! estimates=latent, tables=1:2:3:4 >> between.shw;
quit;
```

Listing A.2: ConQuest input file for the W_{Dif}-MIRT Rasch Model (Data Example A).

```
datafile DataExampleA.dat;
format id 1-5 responses 6-65;
codes 0,1;
set update=yes,warnings=no;
score (0,1) (0,1) (   ) ! items(1-30);  /* Items Yi*/
score (0,1) (0,1) (0,1) ! items(31-60); /* Response Indicators Di*/
model items;
export parameters       >> withinres.prm;
export covariance       >> withinres.cov;
/* Starting estimation & call for item fit (In-/Outfit) */
estimate, fit=yes;
/* Person parameter estimates */
show cases ! estimates = latent >> withinres.lat; /* EAPs & PVs   */
show cases ! estimates = mle    >> withinres.mle; /* ML estimates */
show cases ! estimates = wle    >> withinres.wle; /* WML estimates*/
/* Output */
show ! estimates=latent, tables=1:2:3:4 >> withinres.shw;
quite;
```

In the Rasch-equivalent W_{Res}model, the item parameters in $\tilde{\boldsymbol{\gamma}}_{\xi}$ are also not fixed to zero or one prior to model estimation. Therefore, the application of this model requires software for two-parameter MIRT models. Listing A.3 shows the M*plus* input file of the Rasch-equivalent W_{Res}model used to analyse Data Example A. In line 9, the parameters $\tilde{\gamma}_{im}$ of $\tilde{\boldsymbol{\gamma}}_{\xi}$ are constrained to be equal using the constraint name `'equal'` placed in parentheses. This is implied by the general restriction $\tilde{\gamma}_{im} = \sum_{l=1}^{P} \gamma_{il} b_{lm}$. In Data Example A, that reduces to $\tilde{\gamma}_i = b_1$ since θ and ξ are unidimensional each and $\gamma_{i1} = 1$ for all $i = 1, \ldots, I$. b_1 is the regression coefficient of $E(\theta \mid \xi) = b_0 + b_1 \xi$. Hence, all elements $\tilde{\gamma}_i$ of $\tilde{\boldsymbol{\gamma}}_{\xi}$ have the same value, which is equal to b_1.

Listing A.3: M*plus* input file of the W_{Res}-MIRT Rasch model (Data Example A).

```
DATA:       FILE IS DataExampleA.dat;
            TYPE IS INDIVIDUAL;
VARIABLE:   NAMES ARE id i1-i30 d1-d30;
            USEVARIABLES ARE i1-i30 d1-d30;
            CATEGORICAL ARE i1-i30 d1-d30;
            MISSING IS all (9);
ANALYSIS:   Estimator=MLR;
MODEL:      XI BY i1-i30@1
                  d1-d30(equal); ! Equality Constraint
            RP BY d1-d30@1;
            [XI@0];       ! Restriction: E(xi) = 0
            [RP@0];       ! Restriction: E(xi) = 0
            XI WITH RP@0; ! Restriction: Cov(xi,RP) = 0
OUTPUT:     ...
```

$\tilde{\theta}$ is defined as the residual $\zeta = \theta - E(\theta \mid \xi)$. The expected value $E(\zeta)$ and the covariance $Cov(\xi, \zeta)$ are always zero. This is considered in the model specification in line 13 of

the input file by setting $Cov(\xi, \tilde{\theta}) = 0$. In line 12, the expected value $E(\tilde{\theta})$ is fixed to zero. Furthermore, the expected values $E(\xi)$ and $E(\tilde{\theta})$ are set equal to zero to identify the measurement model of ξ. All thresholds are freely estimated by default in M*plus*.

Two-parameter MIRT models: The 2PL-BMIRT model The two-parameter MIRT models for nonignorable missing data were also applied to Data Example A. The M*plus* input file of the 2PL-BMIRT model is given by Listing A.4. The model was identified by fixing the scale of the latent variables with $Var(\xi) = Var(\theta) = 1$ (line 10) and $E(\xi) = E(\theta) = 0$ (line 11).

Listing A.4: M*plus* input file of the 2PL-BMIRT model (Data Example A).

```
1  DATA:       FILE IS DataExampleA.dat;
2              TYPE IS INDIVIDUAL;
3  VARIABLE:   NAMES ARE id i1-i30 d1-d30;
4              USEVARIABLES ARE i1-i30 d1-d30;
5              CATEGORICAL ARE i1-i30 d1-d30;
6              MISSING IS all (9);
7  ANALYSIS:   Estimator=MLR;
8  MODEL:      XI BY i1* i2-i30; ! Item discrimination
9              RP BY d1* d2-d30; ! Item discrimination
10             ! Model identification
11             XI@1 RP@1;        ! Var(xi) = Var(theta) = 0
12             [XI@0 RP@0];      ! E(xi) = E(theta) = 0
13 OUTPUT:     ...
```

The 2PL-W$_{Dif}$MIRT model The M*plus* input file of the 2PL-W$_{Dif}$MIRT model is given by Listing A.5. The model was identified by the restriction $Var(\xi) = 1$ (line 14) and $E(\xi) = E(\theta) = 0$ (line 15). The variance $Var(\theta^*)$, however, was freely estimated. The reason is that $Var(\theta^*) = Var(\theta - \xi)$. The variance of a difference variable is $Var(\theta - \xi) = Var(\xi) + Var(\theta) - 2 \cdot Cov(\xi, \theta)$. Since $Var(\xi) = 1$ due to model identification, $Var(\theta^*)$ needs to be freely estimated. Therefore, the discrimination parameter γ_i was fixed to be equal to one (line 11). The restriction $\gamma_i^* = 1$ (line 9) is not due to identification but follows from the equality constraint $\gamma_{im}^* = \sum_{i=1}^{P} \gamma_{il}$. Since ξ and θ are undimensional each, that is $\gamma_i^* = \gamma_i$. Hence, the equality $\gamma_i^* = 1$ is implied by the restriction $\gamma_i = 1$.

The 2PL-W$_{Res}$MIRT model The M*plus* input file of the 2PL-W$_{Res}$MIRT model is given by Listing A.6. The model was identified by the restriction $Var(\xi) = 1$ (line 14) and $E(\xi) = E(\tilde{\theta}) = 0$ (line 15). Furthermore, the variance $Var(\tilde{\theta})$ was fixed to one for reasons of model identification. Since $Var(\tilde{\theta}) = Var(\zeta)$, with $\zeta = \theta - E(\theta|\xi)$, the variance of θ is implicitly affected. This affects the parameters γ_i and $\tilde{\gamma}_i$ respectively. However, the

item parameters of the measurement model of ξ as well as the construction and the metric of ξ remain unaffected. From the derivation of the 2PL-W_{Res}MIRT model follows that $\tilde{\gamma}_{im} = \sum_{i=1}^{P} \gamma_{il} b_{lm}$.

Listing A.5: M*plus* input file of the 2PL-W_{Dif}MIRT model (Data Example A).

```
1  DATA:       FILE IS DataExampleA.dat;
2              TYPE IS INDIVIDUAL;
3  VARIABLE:   NAMES ARE id i1-i30 d1-d30;
4              USEVARIABLES ARE i1-i30 d1-d30;
5              CATEGORICAL ARE i1-i30 d1-d30;
6              MISSING IS all (9);
7  ANALYSIS:   Estimator=MLR;
8  MODEL:      XI BY i1* i2-i30     ! Item discrimination
9                    d1@1
10                   d2-d30(a2-a30);! Equality constraints
11             RP BY d1@1           ! Model identification
12                   d2-d30(a2-a30);! Equality constraints
13             ! Model identification
14             XI@1;              ! Var(xi) = 0
15             [XI@0 RP@0];       ! E(xi) = E(theta*) = 0
16 OUTPUT:     ...
```

Since both latent variables ξ and θ are unidimensional, the constraint simplifies to $\tilde{\gamma}_i = \gamma_i b_1$. Again, b_1 is the regression coefficient of $E(\theta|\xi) = b_0 + b_1\xi$. This coefficient is implicitly specified as an additional parameter denoted by `RegC` (line 18) in the model constraint section (lines 17 - 48). The constraints with respect to each parameter $\tilde{\gamma}_i$ of $\tilde{\boldsymbol{\gamma}}$ is specified in the lines 19 - 48 of Listing A.6. Since $Cov(\tilde{\theta}, \xi) = Cov(\zeta, \xi) = 0$, by definition the covariance is fixed to be zero in line 16.

Listing A.6: M*plus* input file of the 2PL-W_{Res}MIRT model (Data Example A).

```
1  DATA:       FILE IS DataExampleA.dat;
2              TYPE IS INDIVIDUAL;
3  VARIABLE:   NAMES ARE id i1-i30 d1-d30;
4              USEVARIABLES ARE i1-i30 d1-d30;
5              CATEGORICAL ARE i1-i30 d1-d30;
6              MISSING IS all (9);
7  ANALYSIS:   Estimator=MLR;
8  MODEL:      XI BY i1* i2-i30     ! Item discrimination
9                    d1*   (a1)     ! Constraint names
10                   d2-d30(a2-a30);! Constraint names
11             RP BY d1*   (g1)     ! Constraint names
12                   d2-d30(g2-g30);! Constraint names
13             ! Model identification
14             XI@1 RP@1;           ! Var(xi) = Var(zeta) = 0
15             [XI@0 RP@0];         ! E(xi) = E(zeta) = 0
16             XI WITH RP@0;        ! Cov(xi,zeta) = 0
17 Model Constraint:
18 new(RegC);
```

```
19 a1  = RegC*g1;
20 a2  = RegC*g2;
21 a3  = RegC*g3;
22 a4  = RegC*g4;
23 a5  = RegC*g5;
24 a6  = RegC*g6;
25 a7  = RegC*g7;
26 a8  = RegC*g8;
27 a9  = RegC*g9;
28 a10 = RegC*g10;
29 a11 = RegC*g11;
30 a12 = RegC*g12;
31 a13 = RegC*g13;
32 a14 = RegC*g14;
33 a15 = RegC*g15;
34 a16 = RegC*g16;
35 a17 = RegC*g17;
36 a18 = RegC*g18;
37 a19 = RegC*g19;
38 a20 = RegC*g20;
39 a21 = RegC*g21;
40 a22 = RegC*g22;
41 a23 = RegC*g23;
42 a24 = RegC*g24;
43 a25 = RegC*g25;
44 a26 = RegC*g26;
45 a27 = RegC*g27;
46 a28 = RegC*g28;
47 a29 = RegC*g29;
48 a30 = RegC*g30;
49 OUTPUT:     ...
```

The relaxed 2PL-W$_{Res}$MIRT model was also applied to Data Example A. In M*plus*, this model can simply be specified by skipping the lines 16 to 47 from Listing A.6. Accordingly, the constraint names are not required in the input file.

The LRM for nonignorable missing data The latent regression model was applied to Data Example A with different functions $f(\boldsymbol{D})$. Here the M*plus* input file of the LRM is shown with the number of completed items (S_D) as the regressor (see Listing A.7).

Listing A.7: M*plus* input file of the LRM (Data Example A).

```
1  DATA:       FILE IS DataExampleA_SD.dat;
2              TYPE IS INDIVIDUAL;
3  VARIABLE:   NAMES ARE id i1-i30 S_D;
4              USEVARIABLES ARE i1-i30 S_D;
5              CATEGORICAL ARE i1-i30;
6              MISSING IS all (9);
7  ANALYSIS:   Estimator=MLR;
8  MODEL:      XI BY i1* i2-i30;    ! Item discrimination
9              ! Latent regression model
10             XI ON S_D (b1);
11             ! For model identification
12             XI   (res);          ! Variance of the latent residual
13             [XI] (int);          ! Intercept
14 Model Constraint: ! for model identification
15 ! Variance of XI is set to one
16 0 = b1**2*XI + res -1;
17 ! Expected value of XI is set to zero
18 0 = int + b1*in1;
19 OUTPUT:     ...
```

The MG-IRT model for nonignorable missing data In M*plus* multiple group IR models can be applied using mixture IRT models with the `KNOWNCLASS`-option. Listing A.8 shows the input file of MG-IRT model for missing responses that was used for Data Example A. The grouping variable `strata` is the stratified response rate. The model was identified by the restriction $E(\xi) = 0$. The group specific means `m1`, `m2`, and `m3`, however, were freely estimated. Since $\boldsymbol{D}$ is informative with respect to the item and person parameters, the means were expected to be different across the groups. The restriction $E(\xi) = 0$ was achieved by setting the mean of the group specific means `m1` to `m3` equal to zero (line 28). In all previous models the variance of the latent variable was fixed to $Var(\xi) = 1$ in order to identify the model. This is difficult in multiple group models. Therefore, the mean of the item discriminations was set to one (lines 29-31). Furthermore, the item discriminations were constrained to be equal across the groups by using group-invariant constraint-names (lines 16, 20, and 24).

Listing A.8: M*plus* input file of the LRM (Data Example A).

```
DATA:       FILE IS ObservedMplusWithStrata.dat;
            TYPE IS INDIVIDUAL;
VARIABLE:       NAMES ARE id i1-i30 d1-d30 strata;
            USEVARIABLES ARE i1-i30;
            CATEGORICAL ARE i1-i30;
            CLASSES = c (3);
            KNOWNCLASS = c (strata=1 strata=2 strata=3);
            MISSING IS all (9);
ANALYSIS:   TYPE IS MIXTURE;
            ALGORITHM = INTEGRATION;
MODEL:      \%OVERALL\%
            XI BY i1-i30;
            [XI*];
            XI;
            \%c#1\%
            XI BY i1-i30 (d1-d30);
            [XI*] (m1);
            XI;
            \%c#2\%
            XI BY i1-i30 (d1-d30);
            [XI*] (m2);
            XI;
            \%c#3\%
            XI BY i1-i30 (d1-d30);
            [XI*] (m3);
            XI;
MODEL CONSTRAINT: ! for identification
0 = 0.338*m1 + 0.361*m2 + 0.301*m3;
0 = (d1+d2+d3+d4+d5+d6+d7+d8+d9+d10+
     d11+d12+d13+d14+d15+d16+d17+d18+d19+d20+
     d21+d22+d23+d24+d25+d26+d27+d28+d29+d30)/30;
OUTPUT: ...
```

Appendix C

In Sections 4.5.3.2 and 4.5.3.3, multidimensional IRT models for nonignorable missing data were further developed to cases with a complex underlying dimensionality. Here in this dissertation, the term complex dimensional structure refers to the fact that either the items Y_i or the response indicators D_i or both are within-item multidimensional. That is, the probabilities $P(Y_i = 1 \mid \boldsymbol{\xi})$ depend on more than one latent dimension ξ_m of $\boldsymbol{\xi}$ and/or some item response propensities $P(D_i = 1 \mid \boldsymbol{\theta})$ depend on more than one latent dimension θ_l of $\boldsymbol{\theta}$. The 2PL-BMIRT -, 2PL-W$_{Dif}$MIRT -, and 2PL-W$_{Res}$MIRT models are equivalent models for nonignorable missing responses. However, model specification especially of 2PL-W$_{Dif}$MIRT - and 2PL-W$_{Res}$MIRT models become more and more difficult with increasing model complexity. In this Appendix, the M*plus* (Muthén & Muthén, 1998 - 2010) input files of the three alternative MIRT models are presented using a simulated data example, denoted as *Data Example C*, with a complex dimensional structure. This data set consists of responses to six items Y_i that constitute the measurement model of a two-dimensional latent ability $\boldsymbol{\xi}$. The latent response propensity $\boldsymbol{\theta}$ underlying the six response indicators D_i is also two-dimensional. Data Example C was simulated according to the path diagram depicted in Figure 4.23. Accordingly, the specified 2PL-BMIRT -, 2PL-W$_{Dif}$MIRT -, and 2PL-W$_{Res}$MIRT models in the following M*plus* input files are graphically represented as path diagrams in the Figures 4.23, 4.24 and 4.25.

Note that the number of items Y_i is very small and not recommended for real applications. However, Data Example C has only been chosen for didactic reasons to show model specification in M*plus* and to demonstrate model equivalence of MIRT models for item nonresponses.

Data Example C The dichotomous items $Y_1, \ldots, Y_6$ constitute the measurement model of $\boldsymbol{\xi} = (\xi_1, \xi_2)$. The items $Y_1 - Y_4$ indicate ξ_1 and Y_2 and $Y_4 - Y_6$ indicate ξ_2. Hence, there is within-item multidimensionality in the items Y_2 and Y_4. The latent response propensity $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is also a two-dimensional latent variable. The response indicators $D_1 - D_3$ constitute the measurement model of θ_1 and $D_2 - D_6$ indicate θ_2. Hence, the items D_2 and D_3 are also within-item multidimensional manifest variables in the measurement model

of $\boldsymbol{\theta}$. All latent dimensions are correlated, implying that the missing data in Data Example C are nonignorable. The true and estimated correlations underlying Data Example C are given in Table 5.3. The positive correlations $Cor(\xi_m, \theta_l)$ imply that the tendency to respond to the items increases with the persons proficiency levels in ξ_1 and ξ_2.

The sample size was $N = 5000$. The true item parameters can be seen in the model equation of the logits in Equation 5.1. This Equation refers to the general model equation given by the Equations 4.79 and 4.80. The four partitions of $\boldsymbol{\Lambda}$ refer to $\boldsymbol{\alpha}$, $\mathbf{0}$, $\boldsymbol{\gamma}_{\xi}$ and $\boldsymbol{\gamma}_{\theta}$. Accordingly, the vector of threshold parameters are partitioned into $\boldsymbol{\beta}$, the vector of item difficulties, and $\boldsymbol{\gamma}_{\mathbf{0}}$, which are the thresholds of the response indicators.

$$\begin{pmatrix} l(Y_1) \\ l(Y_2) \\ l(Y_3) \\ l(Y_4) \\ l(Y_5) \\ l(Y_6) \\ \hline l(D_1) \\ l(D_2) \\ l(D_3) \\ l(D_4) \\ l(D_5) \\ l(D_6) \end{pmatrix} = \left(\begin{array}{cc|cc} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.6 & 0.0 & 0.0 \\ 1.2 & 0.0 & 0.0 & 0.0 \\ 0.6 & 0.4 & 0.0 & 0.0 \\ 0.0 & 1.4 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ \hline 0.0 & 0.0 & 2.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.4 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 1.2 \\ 0.0 & 0.0 & 0.0 & 2.0 \end{array}\right) \begin{pmatrix} \xi_1 \\ \xi_2 \\ \theta_1 \\ \theta_2 \end{pmatrix} - \begin{pmatrix} -2.2 \\ -1.0 \\ 0.0 \\ 0.5 \\ 1.0 \\ 1.5 \\ \hline -1.8 \\ -0.8 \\ -1.3 \\ 0.7 \\ -0.8 \\ 1.2 \end{pmatrix} \tag{5.1}$$

The overall proportion of missing data was 40.7%. The proportion of missing responses per item ranged between 23.3% and 66.1%[5].

The item means $\bar{y}_i$ and $\bar{y}_{i;obs}$ of the complete data and the observed data with missing data can be found in columns two and three of Table 5.4. Due to systematic item nonresponses depending on the latent ability, the item means of the observed data are slightly positively biased, whereas estimated item difficulties $\hat{\beta}_i$ are negatively biased if item nonresponses are ignored. In contrast, the estimated item difficulties of the three MIRT models and the LRM are nearly unbiased.

Table 5.5 shows the true and estimated item discriminations of Data Example C. On average, the item discriminations were slightly underestimated when missing responses are ignored. A small positive bias can be found in discrimination estimates of the MIRT models and the LRM. In Section 3.2.3 it was demonstrated that discrimination parameter estimates are not systematically biased. Insofar, the small biases in the estimated dis-

[5]The proportions of missing responses in the items Y_1 to Y_6 were 25.4%, 31.1%, 23.3%, 63.5%, 34,9%, and 66.1%.

Table 5.3: True and Estimated Correlations of Latent Variables Underlying Data Example C.

	True correlations			
	ξ_1	ξ_2	θ_1	θ_2
ξ_1	1.000			
ξ_2	0.400	1.000		
θ_1	0.800	0.600	1.000	
θ_2	0.600	0.800	0.500	1.000
	Estimated correlations 2PL-BMIRT model			
	ξ_1	ξ_2	θ_1	θ_2
ξ_1	1.000			
ξ_2	0.518	1.000		
θ_1	0.832	0.552	1.000	
θ_2	0.593	0.831	0.447	1.000
	Estimated correlations 2PL-W$_{Dif}$MIRT model			
	ξ_1	ξ_2	θ_1^*	θ_2^*
ξ_1	1.000			
ξ_2	0.521	1.000		
θ_1^*	0.259	-0.411	1.000	
θ_2^*	-0.356	0.049	-0.839	1.000
	Estimated correlations 2PL-W$_{Res}$MIRT model			
	ξ_1	ξ_2	$\tilde{\theta}_1$	$\tilde{\theta}_2$
ξ_1	1.000			
ξ_2	0.510	1.000		
$\tilde{\theta}_1$	-	-	1.000	
$\tilde{\theta}_2$	-	-	-0.479	1.000

crimination parameters should not be interpreted cautiously. Note the similarity of the estimates $\hat{\alpha}_{im}$ of the three MIRT models, which underlines model equivalence.

Table 5.4: Item Means, True and Estimated Item Difficulties for Data Example C.

	Item means		Item difficulties					
Item	$\bar{y}_i$	$\bar{y}_{i;obs}$	True	Ignore	2PL-BMIRT	2PL-W$_{Dif}$MIRT	2PL-W$_{Res}$MIRT	LRM
1	0.870	0.901	-2.2	-2.568	-2.306	-2.290	-2.308	-2.302
2	0.705	0.738	-1.0	-1.100	-1.002	-0.992	-1.002	-0.995
3	0.495	0.522	0.0	-0.083	0.063	0.080	0.065	0.085
4	0.397	0.473	0.5	0.234	0.565	0.561	0.563	0.565
5	0.322	0.384	1.0	0.759	1.075	1.065	1.078	1.080
6	0.225	0.325	1.5	0.939	1.429	1.428	1.424	1.439
Mean bias	<0.001	0.056	-	-0.230	0.004	0.009	0.003	0.012

Table 5.5: True and Estimated Item Discriminations for Data Example C.

		Estimated item discriminations				
Parameter	True	Ignore	2PL-BMIRT	2PL-W$_{Dif}$MIRT	2PL-W$_{Res}$MIRT	LRM
α_{11}	1.0	1.102	1.369	1.311	1.357	1.295
α_{21}	0.5	0.360	0.506	0.501	0.504	0.483
α_{31}	1.2	1.296	1.210	1.234	1.225	1.277
α_{41}	0.6	0.560	0.510	0.505	0.518	0.535
α_{22}	0.6	0.420	0.445	0.454	0.452	0.469
α_{42}	0.4	0.379	0.537	0.535	0.533	0.533
α_{52}	1.4	1.655	1.518	1.497	1.530	1.527
α_{62}	1.0	0.731	0.919	0.918	0.918	0.934
Mean bias	-	-0.025	0.039	0.032	0.042	0.044

Application of IRT models for item nonresponses to Data Example C Five IRT models were applied to Data Example C: (a) the two-dimensional IRT model based on $\boldsymbol{Y}$ that ignores missing data, (b) the 2PL-BMIRT model based on $(\boldsymbol{Y}, \boldsymbol{D})$, (c) the 2PL-W$_{Dif}$MIRT model based on $(\boldsymbol{Y}, \boldsymbol{D})$, (d) the 2PL-W$_{Res}$MIRT model based on $(\boldsymbol{Y}, \boldsymbol{D})$, and (e) the latent regression model with the two latent regressions $E(\xi_1 \mid \hat{\theta}_1, \hat{\theta}_2)$ and $E(\xi_2 \mid \hat{\theta}_1, \hat{\theta}_2)$. The model specifications of the different models in M*plus* are given in the Listings A.9 - A.12. According to the M*plus* syntax rules, comments start with „!“. The latent dimensions θ_l, θ_l^* or $\tilde{\theta}_l$ are denoted by `'rp1'` and `'rp2'` respectively.

2PL-BMIRT model

The 2PL-BMIRT model can easily be specified in M*plus* (see Listing A.9). No constraints are required with respect to item discrimination parameters in $\boldsymbol{\gamma}_\xi$ and $\boldsymbol{\gamma}_\theta$. In real applications, the difficulty is to find the dimensional structure underlying $\boldsymbol{D}$ prior to the application of the model. The 2PL-BMIRT model can be identified in different ways. Here the means and the variances of all latent variables were fixed to zero and one (lines 8-11 of Listing A.9). Hence, $E(\xi_m) = E(\theta_l) = 0$ and $Var(\xi_m) = Var(\theta_l) = 1$, with $l \in \{1, 2\}$ and $m \in \{1, 2\}$. All discrimination parameters and item difficulties were freely estimated. Alternatively, at least one item discrimination per latent dimension could be fixed and the variances could be freely estimated. Similarly, the expected values of the latent dimensions could be estimated if at least one threshold of a manifest variable that indicates a latent dimension is fixed. Note that model identification can become more intricate in cases of within-item multidimensionality.

Listing A.9: M*plus* input file of the 2PL-BMIRT model (Data Example C).

```
1  DATA:        FILE IS DataExampleC.dat;
2               TYPE IS INDIVIDUAL;
3  VARIABLE:    NAMES ARE i1-i6 d1-d6;
4               USEVARIABLES ARE i1-i6 d1-d6;
5               CATEGORICAL ARE i1-i6 d1-d6;
6  ANALYSIS:    Estimator=MLR;
7  MODEL:       ! Item discrimination parameters
8               xi1 BY i1* i2-i4;
9               xi2 BY i2* i4-i6;
10              rp1 BY d1* d2 d3;
11              rp2 BY d2* d3-d6;
12              ! Model identification
13              xi1@1 xi2@1 rp1@1 rp2@1;
14              [xi1@0 xi2@0 rp1@0 rp2@0];
15 OUTPUT:      ...
```

2PL-W$_{Res}$MIRT model

In Listing A.10 the M*plus* input file of the 2PL-W$_{Res}$MIRT model is shown. There are two types of constraints that need to be imposed in this model to ensure correct model specification and model identification: (a) The item discrimination parameters of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{\xi}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$ are constrained to be $\tilde{\gamma}_{im} = \sum_{i=1}^{P} \gamma_{il} b_{lm}$, and (b) the latent covariance $Cov(\xi_m, \tilde{\theta}_l)$ needs to be fixed to zero since $\tilde{\theta}_l$ is defined as the latent residual ζ_l of the regression $E(\theta_l \,|\, \boldsymbol{\xi})$. The constraint estimation of the item discrimination parameters are specified in M*plus* using constraint names `gx11` to `gx62` and `gt11` to `gt62` in lines 8-29. The constraints with respect to γ^*_{im} require the regression coefficients of the latent regressions $E(\theta_1 \,|\, \boldsymbol{\xi}) = b_{10} + b_{11}\xi_1 + b_{12}\xi_2$ and $E(\theta_2 \,|\, \boldsymbol{\xi}) = b_{20} + b_{21}\xi_1 + b_{22}\xi_2$. The four regression coefficients are specified as additional parameters in line 41. Therefore, the latent regression needs not to be specified explicitly in the model command. The constraint with respect to each of the 12 parameters γ^*_{im} are specified in the lines 42-53. The four covariances $Cov(\xi_m, \tilde{\theta}_l)$ are set to zero in lines 35-38.

Listing A.10: M*plus* input file of the 2PL-W$_{Res}$MIRT model (Data Example C).

```
1  DATA:        FILE IS DataExampleC.dat;
2               TYPE IS INDIVIDUAL;
3  VARIABLE:    NAMES ARE i1-i6 d1-d6;
4               USEVARIABLES ARE i1-i6 d1-d6;
5               CATEGORICAL ARE i1-i6 d1-d6;
6  ANALYSIS:    Estimator=MLR;
7  MODEL:     ! Item discrimination parameters
8             xi1 BY i1* i2-i4
9                     d1(gx11)
10                    d2(gx21)
11                    d3(gx31)
12                    d4(gx41)
13                    d5(gx51)
14                    d6(gx61);
15              xi2 BY i2* i4-i6
16                    d1(gx12)
17                    d2(gx22)
18                    d3(gx32)
19                    d4(gx42)
20                    d5(gx52)
21                    d6(gx62);
22              rp1 BY d1*(gt11)
23                    d2(gt21)
24                    d3(gt31);
25              rp2 BY d2*(gt22)
26                    d3(gt32)
27                    d4(gt42)
28                    d5(gt52)
29                    d6(gt62);
30
```

```
31              ! Model Identification
32              xi1@1 xi2@1 rp1@1 rp2@1;
33              [xi1@0 xi2@0 rp1@0 rp2@0];
34
35              xi1 WITH rp1@0;
36              xi1 WITH rp2@0;
37              xi2 WITH rp1@0;
38              xi2 WITH rp2@0;
39
40  MODEL constraint:
41  new(b11 b12 b21 b22);
42  gx11  = gt11*b11;
43  gx21  = gt21*b11 + gt22*b21;
44  gx31  = gt31*b11 + gt32*b21;
45  gx41  =            gt42*b21;
46  gx51  =            gt52*b21;
47  gx61  =            gt62*b21;
48  gx12  = gt11*b12;
49  gx22  = gt21*b12 + gt22*b22;
50  gx32  = gt31*b12 + gt32*b22;
51  gx42  =            gt42*b22;
52  gx52  =            gt52*b22;
53  gx62  =            gt62*b22;
54
55  OUTPUT:     ...
```

2PL-W$_{Dif}$MIRT model

The specification of the 2PL-W$_{Dif}$MIRT model in M*plus* is shown in Listing A.11. As in the case of the 2PL-W$_{Res}$MIRT model, constraint parameter estimation is required. In particular, the discrimination parameters γ^*_{im} and γ_{il} of $\boldsymbol{\gamma}^*_{\boldsymbol{\xi}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\theta}}$ respectively. In lines 35 - 42 of Listing A.10, each element of $\boldsymbol{\gamma}^*_{\boldsymbol{\xi}}$ is constraint to be $\gamma^*_{im} = \sum_{i=1}^{P} \gamma_{il}$. Model identification is given by $E(\xi_1) = E(\xi_2) = E(\theta^*_1) = E(\theta^*_2) = 0$ and $Var(\xi_1) = Var(\xi_2) = 1$. The variances $Var(\theta^*_1)$ and $Var(\theta^*_2)$ were not fixed since both dimensions θ^*_l are defined as latent difference variables $\theta_l - (\xi_1 + \xi_2)$. Hence, the variances are $Var(\theta^*_l) = Var(\theta_l) + \sum_{m=1}^{2}[Var(\xi_m) - 2Cov(\theta, \xi_m)] + 2Cov(\xi_1, \xi_2)$, with θ_l the latent response propensity as defined in the 2PL-BMIRT model. The restriction $Var(\xi_1) = Var(\xi_2) = 1$ for identification of the measurement model of $\boldsymbol{\xi}$ contradicts with a fixed variance of θ^*_l. Therefore, the discrimination parameters γ_{11} and γ_{62} were alternatively fixed to one to identify the measurement model of $\boldsymbol{\theta}^*$ (see lines 21 and 28 of Listing A.10). Accordingly, the parameters γ^*_{11}, γ^*_{61}, γ^*_{12}, and γ^*_{62} were fixed to one (lines 8, 13, 15, and 20). This is not an additional restriction but follows directly from the constraints with respect to the parameters γ^*_{im} of $\boldsymbol{\gamma}^*_{\boldsymbol{\xi}}$ derived in the 2PL-W$_{Dif}$MIRT model (see above).

Listing A.11: M*plus* input file of the 2PL-W$_{Dif}$MIRT model (Data Example C).

```
DATA:       FILE IS DataExampleC.dat;
            TYPE IS INDIVIDUAL;
VARIABLE:   NAMES ARE i1-i6 d1-d6;
            USEVARIABLES ARE i1-i6 d1-d6;
            CATEGORICAL ARE i1-i6 d1-d6;
ANALYSIS:   Estimator=MLR;
MODEL:      ! Item discrimination parameters
            xi1 BY i1* i2-i4
                   1@1
                   d2 (gx21)
                   d3 (gx31)
                   d4 (gx41)
                   d5 (gx51)
                   d6@1;
            xi2 BY i2* i4-i6
                   d1@1
                   d2 (gx22)
                   d3 (gx32)
                   d4 (gx42)
                   d5 (gx52)
                   d6@1;
            rp1 BY d1@1       ! Model identification
                   d2 (gt21)
                   d3 (gt31);
            rp2 BY d2*(gt22)
                   d3 (gt32)
                   d4 (gt42)
                   d5 (gt52)
                   d6@1;      ! Model identification

            ! Model identification
            xi1@1 xi2@1;
            [xi1@0 xi2@0 rp1@0 rp2@0];

MODEL constraint:
gx21 = gt21 + gt22;
gx31 = gt31 + gt32;
gx41 = gt42;
gx51 = gt52;
gx22 = gt21 + gt22;
gx32 = gt31 + gt32;
gx42 = gt42;
gx52 = gt52;

OUTPUT:     ...
```

Latent Regression Model

The LRM for missing responses consists of two parts that need to be specified in the M*plus* input file: (a) the measurement model of $\boldsymbol{\xi}$ and (b) the latent regression model with

$E[\xi_1 \mid f(\boldsymbol{D})]$ and $E[\xi_2 \mid f(\boldsymbol{D})]$. The measurement model of $\boldsymbol{\xi}$ is described in lines 7-10 of Listing A.12. The EAP estimates of the latent response propensities θ_1 and θ_2 were used as independent variables in the latent regression model and were generated in a previous step using a two-dimensional two-parameter IRT model for the response indicators $D_1, \ldots, D_6$. The measurement model of $\boldsymbol{\theta}$ was specified according to the true data-generating model. Note that the appropriate model for $\boldsymbol{D}$ needs to be explored in real applications to ensure bias correction (see Section 4.5.3.4). In Listing A.12, the EAP estimates of θ_1 and θ_2 are denoted by `eap1` and `eap2`. The two latent regressions $E[\xi_1 \mid \hat{\theta}_1, \hat{\theta}_2] = b_{10} + b_{11}\hat{\theta}_1 + b_{12}\hat{\theta}_2$ and $E[\xi_2 \mid \hat{\theta}_1, \hat{\theta}_2] = b_{20} + b_{21}\hat{\theta}_1 + b_{22}\hat{\theta}_2$ are specified in lines 18 and 19.

To compare the item and person parameter estimates across different IRT models, a common metric of the latent variables ξ_1 and ξ_2 needs to be established in all IRT models. The simple IRT model that ignores missing responses as well as the MIRT models for nonignorable missing data were identified by setting $E(\xi_1) = E(\xi_2) = 0$ and $Var(\xi_1) = Var(\xi_2) = 1$. In M*plus*, the variance and the expected value of dependent variables in regression models cannot directly be fixed to certain values. Instead, the specification of nonlinear constraints are required. The variance of ξ_m (with $m =\in \{1, 2\}$) in Data Example C is $Var(\xi_m) = b_{l1}^2 Var(\hat{\theta}_1) + b_{l2}^2 Var(\hat{\theta}_2) + 2b_{l1}b_{l2}Cov(\hat{\theta}_1, \hat{\theta}_2) + Var(\zeta_m)$. If the variance is fixed to one, then $0 = b_{l1}^2 Var(\hat{\theta}_1) + b_{l2}^2 Var(\hat{\theta}_2) + 2b_{l1}b_{l2}Cov(\hat{\theta}_1, \hat{\theta}_2) + Var(\zeta_m) - 1$. This expression can be used as a nonlinear constraint in M*plus* (lines 22-23). Similarly, the expected values are $E(\xi_m) = b_{m0} + b_{m1}E(\hat{\theta}_1) + b_{m2}E(\hat{\theta}_2)$. Therefore, the left sight of this equation was set to zero in lines 25 and 26. In M*plus*, the specification of nonlinear constraints requires constraint names, which are placed in parentheses in Listing A.12.

Listing A.12: M*plus* input file of the LRM (Data Example C).

```
1  DATA:       FILE IS DataExampleC.dat;
2              TYPE IS INDIVIDUAL;
3  VARIABLE:   NAMES ARE i1-i6 d1-d6 eap1 eap2;
4              USEVARIABLES ARE i1-i6 eap1-eap2;
5              CATEGORICAL ARE i1-i6 eap1-eap2;
6  ANALYSIS:   Estimator=MLR;
7  MODEL:      ! Item discrimination parameters
8              xi1 BY i1* i2-i4;
9              xi2 BY i2* i4-i6;
10             ! Latent variables
11             [xi1 xi2](al1-al2);     ! Intercepts
12             xi1 xi2(res1-res2);     ! Residual variances
13             ! Independent variables of the LRM
14             eap1-eap2 (v1-v2);      ! Variances
15             [eap1-eap2] (in1-in2); ! Means
16             eap1 WITH eap2 (cov);   ! Covariance
17             ! Latent Regression model
18             xi1 ON eap1 eap2 (g1-g2);
```

```
            xi2 ON eap1 eap2 (b1-b2);
Model Constraint: ! for model identification
! Variances of both latent dimensions are set to one
0 = g1**2*v1 + g2**2*v2 + 2*g1*g2*cov + res1 -1;
0 = b1**2*v1 + b2**2*v2 + 2*b1*b2*cov + res2 -1;
! Expected values of both latent dimensions are set to zero
0 = al1 + g1*in1 + g2*in2;
0 = al2 + b1*in1 + b2*in2;
OUTPUT:     ...
```

Model Fit (Data Example C)

The 2PL-BMIRT -, the 2PL-W$_{Dif}$MIRT -, the 2PL-W$_{Res}$MIRT model, and the LRM as specified in the Listings A.9 - A.12 were applied to Data Example C. The estimates $\hat{\alpha}_{im}$ and $\hat{\beta}_i$ are given in the Tables 5.4 and 5.5. In Table 5.6, different goodness-of-fit statistics and the number of estimated parameters (n_{par}) in the respective model are shown. The three alternative MIRT models were nearly identical in terms of model fit. The item and person parameter estimates of the LRM are also close to that of the MIRT models, which reflects model equivalence in the construction of the latent variables and the reduction of bias. However, the number of parameters is substantially lower in the LRM. The model fit indices of the LRM are quite different from that of the MIRT models and cannot be compared. Recall that the LRM and the MIRT models are not equivalent in terms of model fit (see Section 4.5.4).

The EAP person parameter estimates of ξ_1 and ξ_2 are shown in Figure 5.1. The correlations are shown within the single scatter plots. The EAPs of the different IRT models for item nonresponses are very close to each other but differ substantially from the EAPs of the model that ignores missing data. The EAPs from the MIRT models and the LRM correlates substantially higher with the true value of ξ_1 and ξ_2 than the EAPs obtained by the IRT model that ignores item nonresponses.

Table 5.6: Goodness-of-fit Indices of (M)IRT models for Nonignorable Missing Responses Applied to Data Example C.

Model	Log-Lik.	n_{par}	AIC	BIC
2PL-BMIRT model	-27357.638	34	54783.276	55004.861
2PL-W$_{Res}$MIRT model	-27357.666	34	54783.331	55004.916
2PL-W$_{Dif}$MIRT model	-27358.933	34	54785.866	55007.451
LRM	-18440.700	24	36929.401	37085.813

Note: n_{par} = Number of estimated parameters.

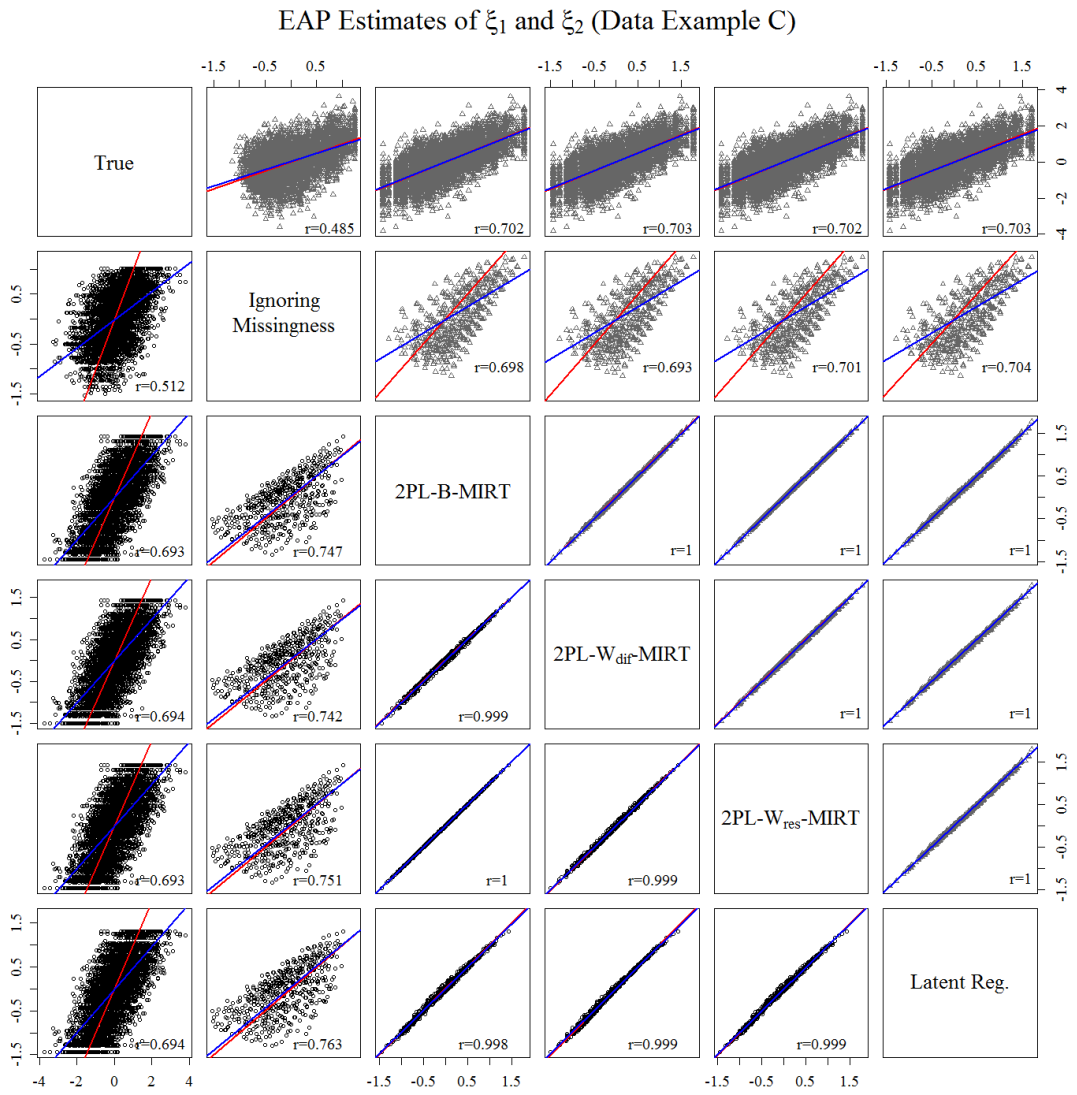

Figure 5.1: Person parameters ξ_1 and corresponding EAP estimates (above diagonal) and person parameters ξ_2 and corresponding EAP estimates (below diagonal) using Data Example C. The red lines indicate the bisectric. The blue lines are regression lines.

Ehrenwörtliche Erklärung

Die Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität in der geltenden Fassung ist mir bekannt.

Ich habe diese Dissertation selbst angefertigt und dabei insbesondere die Hilfe eines Promotionsberaters nicht in Anspruch genommen. Alle von mir benutzten Quellen und Hilfsmittel habe ich kenntlich gemacht und an den entsprechenden Stellen angegeben.

Christiane Fiege, Anna-Lena Dicke und Jessika Golle haben unentgeltlich Vorabversionen einzelner Teile des Manuskriptes gelesen und mich auf Fehler und Inkonsistenzen aufmerksam gemacht. Marlena Itz hat Vorabversionen des Manuskriptes gelesen und mich entgeltlich auf Fehler in den englischen Formulierungen hingewiesen.

Darüber hinaus haben Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe diese Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

Ich habe weder die gleiche noch eine in wesentlichen Teilen ähnliche noch eine andere Arbeit bei einer anderen Hochschule oder Fakultät als Dissertation eingereicht.

Ich versichere, dass die oben gemachten Angaben nach meinem besten Wissen der Wahrheit entsprechen und ich nichts verschwiegen habe.

Jena, den _____ _____

Norman Rose

Lebenslauf

Norman Rose

Geboren am 16.04.1974 in Jena

Familienstand: ledig

1980 - 1990	Polytechnische Oberschule Otto-Grotewohl, Jena (Realschulabschluss)
1990 - 1993	Berufsaubildung zum staatlich anerkannten Krankenpfleger an der Friedrich-Schiller-Universität Jena
1993 - 1995	Tätigkeit als Krankenpfleger in der Klinik für Anästhesie und Intensivmedizin der Friedrich-Schiller-Universität Jena
4/1995 - 03/1996	Zivildienst im staatlichen regionalen Förderzentrum mit dem Förderschwerpunkt geistige Entwicklung Kastanienschule
1996 - 1997	Tätigkeit als Krankenpfleger in der Klinik für Anästhesie und Intensivmedizin der Friedrich-Schiller-Universität Jena
1998 - 2001	Tätigkeit als Krankenpfleger im chirurgischen Operationssaal mit berufsbegleitender Qualifizierung zum Fachkrankenpfleger für den Operationsdienst (DKG)
2001 - 2002	Fachoberschule im Bereich Sozialwesen, an der Staatlichen Berufsbildenden Schule für Gesundheit und Soziales in Jena
2002 - 2007	Studium der Psychologie an der Friedrich-Schiller-Universität Jena
2007 - 2011	Wissenschaftlicher Mitarbeiter und Promotionsstudent am Institut für Psychologie der Friedrich-Schiller-Universität Jena, Lehrstuhl für Methodenlehre und Evaluationsforschung
ab 9/2011	Wissenschaftlicher Mitarbeiter an der Eberhard Karls Universität Tübingen, Abteilung für Empirische Bildungsforschung und Pädagogische Psychologie

Jena, den _____ _____

Norman Rose