Isn’t Something Missing?

Latent Variable Models Accounting for Item Nonresponse

Dissertation

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

im Fachbereich Erziehungswissenschaften und Psychologie

der Freien Universität Berlin

vorgelegt von

Dipl.-Psych. Carmen Köhler

Bamberg, 30.09.2015

Erstgutachter: Prof. Dr. Steffi Pohl (Freie Universität Berlin)

Zweitgutachter: Prof. Dr. Claus H. Carstensen (Otto-Friedrich-Universität Bamberg)

Tag der Disputation: 26.01.2016

Zusammenfassung

Die Nichtbeantwortung einzelner Aufgaben bei der Bearbeitung von Kompetenztests stellt eine Bedrohung für die valide und reliable Messung von Kompetenzen dar, besonders wenn die fehlenden Werte systematisch auftreten und mit der unbeobachteten Antwort zusammenhängen. Dies ist im Kontext großangelegter Bildungsstudien üblicherweise der Fall, da hier die Nichtbeantwortung von Items häufig mit der Fähigkeit der Probanden zusammenhängt. Wissenschaftler haben neue Ansätze entwickelt, welche die Abhängigkeit zwischen Kompetenz und Nichtbeantwortung von Items berücksichtigen. Hierfür wird der Prozess, der zu fehlenden Werten führt, modelliert und in das Messmodell zur Fähigkeitsschätzung mitaufgenommen. Diese vielversprechenden *modellbasierten Ansätze* sind den allgemein gebräuchlichen Ansätzen zum Umgang mit fehlenden Werten—welche den Zusammenhang zwischen Fähigkeit und Nichtbeantwortung nicht berücksichtigen—möglicherweise überlegen. Bislang wurden die Ansätze nur selten auf Anwendbarkeit und Performanz bei der Skalierung von Kompetenztests großangelegter Bildungsstudien geprüft.

Die vorliegende Dissertationsarbeit schließt die Lücke zwischen den theoretisch postulierten Modellen und deren möglicher Implementierung im Kontext großangelegter Bildungsstudien. Ziele der Arbeit waren (1) die Anwendbarkeit modellbasierter Ansätze in Bezug auf Kompetenztests zu prüfen und (2) zu evaluieren ob und unter welchen Bedingungen die Ansätze den allgemein gebräuchlichen Ansätzen vorzuziehen sind. Diesem Zweck dienten drei Forschungsstudien. Studie 1 prüfte, ob die Annahmen modellbasierter Ansätze in Kompetenztestdaten Bestand haben und sich Verletzungen dieser Annahmen auf individuelle Personenparameterschätzer auswirken. Studie 2 untersuchte Merkmale des Antwortverhaltens von Probanden, wie beispielsweise ob die Tendenzen zur Nichtbeantwortung über verschiedene Tests hinweg ähnlich ausfallen und ob sie mit weiteren Personencharakteristika zusammenhängen. Studie 3 testete die Performanz der modellbasierten Ansätze im Vergleich zu allgemein gebräuchlichen Ansätzen.

Ergebnisse zeigen, dass modellbasierte Ansätze durchaus auf Kompetenztestdaten großangelegter Bildungsstudien anwendbar sind, wobei leichte Erweiterungen der Modelle zu einer genaueren Parameterschätzung führen. Des Weiteren können die Tendenzen zur Nichtbeantwortung als personenspezifische Attribute angesehen werden, welche über verschiedene Kompetenztests relativ stabil sind und auch stabile Zusammenhänge zu anderen Persönlichkeitscharakteristika aufweisen. Die Befunde der dritten Studie bestätigen die Überlegenheit der modellbasierten Ansätze gegenüber allgemein gebräuchlichen Methoden, wobei ein Modell, welches fehlende Werte lediglich ignoriert, auch akzeptable Resultate liefert.

Modellbasierte Ansätze weisen im Vergleich zu allgemein gebräuchlichen Ansätzen einige Vorteile auf. In Anbetracht ihrer Komplexität sollten jedoch die Vor- und Nachteile verschiedener Skalierungsmethoden gegeneinander abgewogen werden. Als wichtige Aspekte gelten hierbei die Komplexität des Modells, Auswirkungen auf das Testverhalten der Probanden und die Genauigkeit bei der Schätzung der Parameter. Vielen großangelegten Bildungsstudien steht definitiv ein Wandel in der Art des Umgangs mit fehlenden Werten bevor. Ob modellbasierte Ansätze die bisherigen Methoden ersetzen ist noch unklar. Fest steht, dass modellbasierte Ansätze zu den fortschrittlichsten Methoden zum Umgang mit fehlenden Werten bei der Skalierung von Kompetenztestdaten gehören.

Summary

Item nonresponse in competence tests pose a threat to a valid and reliable competence measurement, especially if the missing values occur systematically and relate to the unobserved response. This is often the case in the context of large-scale assessments, where the failure to respond to an item relates to examinee ability. Researchers developed methods that consider the dependency between ability and item nonresponse by incorporating a model for the process that causes missing values into the measurement model for ability. These *model-based approaches* seem very promising and might prove superior to common missing data approaches, which typically fail at taking the dependency between ability and nonresponse into account. Up to this point, the approaches have barely been investigated in terms of applicability and performance with regard to the scaling of competence tests in large-scale assessments.

The current dissertation bridges the gap between these theoretically postulated models and their possible implementation in the context of large-scale assessments. It aims at (1) testing the applicability of model-based approaches to competence test data, and (2) evaluating whether and under what missing data conditions these approaches are superior to common missing data approaches. Three research studies were conducted for this purpose. Study 1 investigated the assumptions of model-based approaches, whether they hold in empirical practice, and how violations to those assumptions affect individual person parameters. Study 2 focused on features of examinees' nonresponse behavior, such as its stability across different competence tests and how it relates to other examinee characteristics. Study 3 examined the performance of model-based approaches compared to other approaches.

Results demonstrate that model-based approaches can be applied to large-scale assessment data, though slight extensions of the models might enhance accuracy in parameter estimates. Further, persons' tendencies not to respond can be considered person-specific attributes, which are relatively constant across different competence tests and also relate to

other stable person characteristics. Findings from the third study confirmed the superiority of the model-based approaches compared to common missing data approaches, although a model that simply ignores missing values also led to acceptable results.

Model-based approaches show serval advantages over common missing data approaches. Considering their complexity, however, the benefits and drawbacks from different methods need to be weighed. Important issues in the debate on an appropriate scaling method concern model complexity, consequences on examinees' test-taking behavior, and precision of parameter estimates. For many large-scale assessments, a change in the missing data treatment is clearly necessary. Whether model-based approaches will replace former methods is yet to be determined. They certainly count amongst the most advanced methods to handle missing values in the scaling of competence tests.

Table of Contents

Chapter 1

Introduction and Theoretical Background

1 Introduction and Theoretical Background

One of the major difficulties in educational and psychological sciences lies in measuring unobservable constructs. Constructs are psychological or social phenomena, which cannot be directly observed, but need to be inferred from manifest indicators (Bortz & Döring, 2006). Large-scale assessment studies aim at providing data that allows comparisons of students' unobservable proficiencies.[1] In competence tests of large-scale assessment studies, the different tasks examinees are presented with—also referred to as items—serve as manifest indicators of the proficiency construct. A problem that arises in these assessments is that, for various reasons, not all examinees respond to all items. The objective for the researcher is that deviations between inferences drawn from these incomplete data and inferences that would have been obtained if the data had been complete are reduced to a minimum (Rubin, 1976; Mislevy & Wu, 1996). Currently, different large-scale assessment studies employ various missing data approaches, though none of them can be considered state-of-the-art (Schafer & Graham, 2002). The commonly applied methods fail to account for missing responses that depend on the unobserved proficiency, which might result in biased parameter estimates from the measurement model for that particular proficiency (Mislevy & Wu, 1988, 1996). For this reasons, model-based approaches were developed, which incorporate information of an examinee's tendency to respond to an item in the measurement model (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Moustaki & O'Muircheartaigh, 2000). These advanced models seem very promising, but so far have not been extensively tested for application in large-scale assessment studies. Furthermore, the models are rather complex, and more parsimonious models might suffice to retrieve unbiased parameter estimates. The present dissertation intends to (1) investigate the properties of the process underlying missing data in large-scale assessments, and (2) aims to find a scaling

[1] In the present dissertation, the terms proficiency, ability, and competence are used interchangeably.

model that adequately takes the missing responses into account. The three research papers conducted within this dissertation centered around these goals, focusing on (a) whether underlying assumptions regarding the model-based approaches hold in empirical data, and whether violations to these assumptions affect person parameter estimates, (b) evaluating possible extensions of the model-based approaches in order to scale several competence domains simultaneously and to allow for the inclusion of covariates that influence the nonresponse behavior, and (c) evaluating which missing data approach should be applied in the scaling of competence tests.

The first chapter (Chapter 1) sets up the broad theoretical framework of the dissertation and introduces the main research topics. It is followed by the three research papers covering those research topics (Chapters 2-4). The dissertation concludes with a discussion chapter (Chapter 5). The first section of Chapter 1 (1.1) revolves around large-scale assessments and points out the relevance of missing data occurring in competence tests. The main focus of the section lies on different missing data types and how they are dealt with in the scaling of competence tests. The second section (1.2) discusses whether the common missing data treatment is actually adequate. It introduces theoretical considerations on how missing values influence and potentially bias parameter estimates, and outlines under which missing data conditions adequate inferences can be drawn from the data. These theoretical concepts are subsequently transferred to the context of large-scale assessments. It is pointed out that the common treatment of missing values in large-scale assessments is incompatible with the theoretical concepts on an adequate missing data treatment. In the third section (1.3), model-based approaches are introduced, which might hold the solution to reconciling theory and practice. The section centers on different types of model-based approaches and discusses current research on their application to real data. It sets the stage for the research questions of the present dissertation, which are outlined in section 1.4. The research topics tie into previous work on model-based approaches, extending investigations on the missing data mechanism in

competence tests and evaluating different missing data approaches. The section describes the three research studies that were conducted for this purpose, embedding them in the overall theoretical framework. The last chapter (Chapter 5) summarizes the main results (5.1), draws conclusions for application and practice (5.2), and debates future research perspectives (5.3).

1.1 Large-Scale Assessments and Missing Data

"Missingness is usually a nuisance,

not the main focus of inquiry"

(Schafer & Graham, 2002, p. 147)

The following section of the chapter introduces educational large-scale assessment studies and the role that missing values play in this context. The focus of large-scale studies is to provide large amounts of representative data on student competencies, which allow for comparisons in education between collective groups (Blossfeld, von Maurice, & Schneider, 2011). The occurrence of missing values is oftentimes considered a nuisance in these assessments (Schafer & Graham, 2002). If missing values are not handled correctly, however, they can threaten the validity and reliability of any analysis conducted with the data (Mislevy & Wu, 1996). The main focus of the present dissertation lies on missing values and how they should be treated so they remain a nuisance and not a conceivable source of bias. Since our core objectives are embedded in the context of competence measurement in large-scale assessments, our investigations on missing data need to be discussed and put into context with the conditions and objectives of large-scale assessment studies. The first section of this dissertation therefore mainly focusses on large-scale assessments in general, and the role that missing values have played in them. It starts by giving a brief overview of the history and purpose of large-scale assessments studies (1.1.1). The main focus of the section lies on describing how various studies deal with missing values (1.1.2). These are not considered best practice by many researchers, and fueled discussion on an adequate missing data treatment.

The last subsection outlines some of the early voiced criticism, and introduces several developments in the missing data research (1.1.3).

1.1.1 History and purpose of large-scale assessments

Proficiency can be considered human capital: It brings advancements in relevant fields such as technology and science, and, in turn, largely contributes to the prosperity of a country (von Davier, Gonzalez, Kirsch, & Yamamoto, 2013). Consequently, the development and acquirement of skills within the population of a country is of political interest. Large-scale assessment studies aim at providing information on skill levels by administering objective standardized tests to a certain representative subset of the targeted population. Such studies are often funded by the government, and exist on the national as well as the international level. On the national level, standardized tests serve as a tool to measure and evaluate educational growth. In the U.S., they are even used to monitor the performance of individuals, schools, and districts (DePascale, 2003). The aim with regard to tests on the international level is to compare the efficiency between different educational systems (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962). Overall, large-scale assessment studies majorly impact educational decisions, and have resulted in numerous educational reforms (Baird, et al., 2011; DePascale, 2003).

The assessments mostly focus on core competence domains such as mathematics, reading, and science. Today, some of the most widely known and influential international large-scale assessment studies are PISA (Programme for International Student Assessment), TIMSS (Third International Mathematics and Science Study), and PIRLS (Progress in International Reading Literacy Study). The first international PISA survey was launched in 2000, assessing reading literacy, mathematical literacy, and scientific literacy of 15-year-olds; TIMSS started in 1994, focusing on mathematic skills of-14 year-olds; the first PIRLS survey took place in 2001, assessing reading literacy of 10-year-olds (Baird, et al., 2011). Besides the international studies, many countries additionally conduct large-scale assessment studies on

the national level, for example, NAEP (National Assessment of Educational Progress) in the U.S., NAP (National Assessment Program) in Australia, or NEPS (National Educational Panel Study) in Germany. NAEP targets the age groups 9, 13, and 17, NAP assesses competencies from students in 6^{th} and 10^{th} grade, and NEPS covers the entire life-span from newborns to adults. These examples illustrate the scope of large-scale assessments, and underline the vast range of the population that is potentially affected by evaluations from large-scale assessment studies.

As already mentioned, the main aim of large-scale studies is to allow for competence comparisons. These include, for example, comparisons between gender groups or ethnicity groups, and comparisons between different school classes, districts, or countries. Studies with longitudinal designs additionally intend to allow for comparisons over time. The tests mentioned are considered low-stakes achievement tests insofar that the scores are only regarded at the aggregate group-level, and test-takers do not receive individual feedback on how they performed (Baumert & Demmrich, 2001). They are thus distinct from standardized high-stakes achievement tests such as college entrance tests (e.g., SAT) or language tests (e.g., TOEFL [Test of English as a Foreign Language]). This differentiation is relevant with regard to missing data treatment, since the type of test naturally plays into the motivation to perform well on the test. Examinee motivation, in turn, affects test-taking strategies and response behavior (Wise & DeMars, 2005). How assumptions about strategies and response behavior influence the missing data treatment is discussed in the following subsection, which first introduces the scaling of competence tests and subsequently discusses different types of missing values and how they are commonly handled in the scaling of low-stakes tests.

1.1.2 Missing values in large-scale assessments

Competence tests in large-scale assessments usually comprise several tasks in each administered competence domain. Since NAEP introduced *Item Response Theory* (IRT) into the psychometric scaling of competencies in the 1980s, IRT models have been the

predominant measurement method in large-scale assessments (von Davier et al., 2013). In IRT, the answers to the items serve as manifest indicators of the underlying latent construct to be measured. The models basically try to describe the entire data set by putting the response to the item, the difficulty of the item, and person ability into relation (Embretson & Reise, 2000). In the most basic IRT model, the Rasch model, the prediction of a correct response is simply described as a function of person ability and item difficulty (Rasch, 1960). IRT models show several psychometric advantages over approaches from classical test theory (CTT), including the comparability of persons even when they were presented with different items (Bock, 1997). This is due to the fact that the estimation of item parameters is independent of which persons responded to the items, and the estimation of person parameters is irrespective of which items the persons were presented with. Based on the observed data, item parameters and person parameters are estimated—parameter estimation can thus simply be based on all observed responses (Embretson & Reise, 2000). Unobserved responses are dropped from the likelihood equation for parameter estimation, and thus do not contribute to these parameter estimates (Lord, 1974).

The fact that unobserved responses do not contribute to the estimation can threaten the accuracy of item and person parameters if the unobserved response holds information beyond what has been captured in the observed data (Lord, 1974; Mislevy & Wu, 1988, 1996). Missing values might occur according to a certain mechanism, and the option of ignoring missing values is truly only applicable for certain types of missing values (see Section 1.2). Before going into detail on when and why missing values are truly ignorable, the following paragraphs first describe original assumptions regarding the occurrence of missing data in large-scale assessments and the corresponding treatments of missing data. These assumptions are reconsidered and reevaluated in Section 1.2.

In large-scale assessments, the treatment of missing values in the scaling depends on the type of missing value. The most prominent types of missing responses are *not-*

administered items, *not-reached items* and *omitted items*.[2] Not-administered items result from examinees being presented with alternate test forms (see, e.g., OECD, 2009). This type of missing response is therefore planned, which reflects the fact that IRT can typically handle nonresponse. As the reason for the nonresponse lies in the setup of the study and the test forms are distributed randomly to the examinees, the inferences on ability are irrespective of which items were administered to which examinee (Mislevy & Wu, 1988, 1996). In large-scale assessments, missing values due to not-administered items are therefore ignored in the scaling, meaning that the item parameter estimation is only based on persons who were given the respective items, and person parameter estimation is only based on those items that were administered to the respective person (Mislevy & Wu, 1988, 1996).

Not-reached items are missing responses toward the end of a test. Although tests in low-stakes assessments are typically not speeded, they are mostly administered with a certain time limit (see, e.g. OECD, 2009; Pohl, Haberkorn, Hardt, & Wiegand, 2012). This keeps some examinees from finishing the test. Usually, all items after the last given valid response are labelled *not-reached* (Beaton, 1987). The most common perception with regard to the occurrence of not-reached items is that the examinee did not attempt the item, and therefore the item—or the missingness thereof—holds no additional information about the underlying construct of interest (Lord, 1974; Mislevy & Wu, 1988, 1996). Ignoring not-reached items thus underlies the assumption that the estimated ability level based on the observed items does not substantially differ from the ability level the examinee would have obtained if he or she had been given ample time to respond to all items (Lord, 1974; Mislevy & Wu, 1988, 1996). Two conditions are necessary for these assumptions to hold, namely that the item positions

[2]The present dissertation focusses solely on item nonresponse (i.e., not all items receive a response, but the examinee generally participates in the test), which is distinguished form unit nonresponse (i.e., the examinee fails to respond entirely).

are random with respect to item difficulty, and that examinees work on the items in the order the items are presented to them, starting from the beginning of the test (Lord, 1974; Mislevy & Wu, 1988, 1996). If the assumptions hold, the inferences on the parameter estimates can solely be based on the items the examinee gave a response to (Lord, 1974; Mislevy & Wu, 1988, 1996). Accordingly, some large-scale assessments such as NAEP and NEPS ignore not-reached items (see Johnson & Allen, 1992; Pohl & Carstensen, 2012), though several assessments use a slightly altered method. PISA, TIMSS, PIRLS, and NAP, for example, employ a two-stage process: (1) The items are ignored in the step of calibrating item parameters; (2) they are treated as incorrect responses when estimating person ability parameters, meaning that each not-reached item receives a score of 0 (Adams & Wu, 2002; Ainley, Fraillon, & Freeman, 2008; Foy, Galia, & Li, 2008; Martin, Gregory, & Stemler, 2000). [3]

The term *omission* or *omitted item* refers to all missing items within the test which are followed by one or several valid responses (Beaton, 1987). The widely accepted assumption with regard to item omissions is that the examinee reached and attempted the item, but failed to respond (Lord, 1974; Mislevy & Wu, 1996). Since the test taker did have sufficient time to appraise the item, he or she apparently could not produce the correct answer, and the omitted item should thus be scored *incorrect*. Accordingly, most studies such as PISA, TIMSS, PIRLS, and NAP treat omitted items as incorrect items (Adams & Wu, 2002; Ainley et al., 2008; Foy et al., 2008; Martin et al., 2000). In NAEP, omitted multiple-choice items are treated as fractionally correct items, meaning that the reciprocal of the number of response

[3]The reason why many large-scale assessments choose to score not-reached items as incorrect responses in the step of generating proficiency estimates is mostly based on fear of abuse of the scaling method rather than on assumptions regarding not-reached items. The concern is that examinees might adapt their answering strategy or be less motivated to respond if there was no penalty for leaving items unanswered (Culbertson, 2011; Lord, 1974).

choices is imputed for all omitted items; items with other response formats are treated as incorrect responses (Johnson & Allen, 1992). In NEPS, omitted items are ignored in the scaling of item and person parameters (Pohl & Carstensen, 2012). The reason why NAEP and NEPS use alternate methods for omitted items lies in criticism with regard to the method and the underlying assumptions of treating missing values as incorrect responses (Johnson & Allen, 1992; Pohl & Carstensen, 2012). The early criticism expressed by Lord (1974) is addressed in the following subsection, and is further discussed in Section 1.2.

1.1.3 Early criticism of commonly employed missing value practices and new developments

To assure valid inferences from analyses with competence data, competencies need to be accurately measured. Lord (1974) was one of the first researchers to proclaim that "omitted items cannot properly be treated as wrong when estimating ability and item parameters." (p. 247). He argues that for multiple-choice items, people would be working against their best interest when they omit an item, since their chance probability for success on an item is the reciprocal of the number of response choices. When treating omissions as incorrect, the probability for a correct score is fixed at 0, whereas it should at least be at chance level. Lord (1974) also argues against fractional correct scoring, however, since the chance level cannot truly be considered a constant, but is different for examinees at different ability levels. He also argues against simply ignoring an omitted item, since the omission carries information about the examinees uncertainty of the correct answer. He proposes to replace omitted items by randomly imputing a 0 or a 1, whereat the probability of imputing a 1 is set equal to the reciprocal of the number of response choices—basically the guessing constant. Note that his method assumes that the chance of success is roughly the guessing constant for all examinees.

This proposition has not been put into practice in large-scale assessments, and Lord's criticism of the scaling of omitted items received relatively little attention. In the following years, the discussion of an accurate missing data treatment was limited to theoretical papers on the ignorability of missing values. These include Rubin's (1976) seminal paper, in which

he outlines under what conditions the missing data mechanism can be ignored. Several years later, Mislevy and Wu (1988, 1996) transferred Rubin's work into the IRT context, discussing the ignorability of several types of missing responses that typically occur in large-scale assessments. Since then, several alternative methods for dealing with missing values were proposed, including imputation techniques (Rubin, 1987), weighting approaches (Robins, Rotnitzky, & Zhao, 1994), and model-based approaches (Glas & Pimentel, 2008; Glynn, Laird, & Rubin, 1986; Heckman, 1976; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999).

Increasingly in the last decade, simulation studies and empirical studies were conducted to establish the consequences of some of the newer missing data techniques as well as the more traditional methods on item and person parameter estimation. Though many of them seem quite promising and most studies established that treating missing values as incorrect responses leads to an underestimation of ability estimates (Culbertson, 2011; de Ayala, Plake, & Impara, 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Pohl, Gräfe, & Rose, 2014; Rose, von Davier, & Xu, 2010), none of these newer approaches have been put into practice in large-scale assessments.

1.2 Dealing with Missing Data

> "Our objective is to find the weakest simple conditions on the process that causes missing data such that it is always appropriate to ignore this process […].
>
> The conditions turn out to be rather intuitive"
>
> (Rubin, 1976, p. 582)

This section describes in detail the previously mentioned theoretical basics on how to appropriately handle missing values, outlining the conditions of ignorability (1.2.1). Subsequently, the proposed conditions for ignorability and the consequences for the scaling of different types of missing values occurring in large-scale assessments are discussed (1.2.2).

1.2.1 Conditions for ignorability of missing data

Rubin (1976) introduced the terms *missing at random* (MAR), referring to the randomness of the missing data, and *observed at random* (OAR), pertaining to the randomness of the observed data. His definition emphasizes the distinction between two sources of influence which the missingness may depend upon. One of them lies in the variable to be measured, the other lies in a mechanism behind the missingness itself. Both may affect the probability of a missing value, thus determining which of the data is observed (Rubin, 1976). His definition of MAR states that for each possible value of the parameter of the missing data process, the probability of the observed missing data conditional on the observed data must be the same for any possible value of the unobserved data. This means that the missing data process can depend on the observed variable, but is the same for each value of the observed variable. His definition of OAR states that for any possible observed value, the conditional distribution responsible for the missingness is the same for each possible value of the parameter of the missing data process and each possible missing data value. Basically this implies that the probability for making an observation does neither depend on the missing data process nor on the observed pattern of the missing data (Schafer & Graham, 2002).

Later, the three distinctions *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR) became more prominent (see, e.g., Holman & Glas, 2005; Mislevy & Wu, 1988, 1996; Schafer & Graham, 2002). MCAR equals Rubin's definition of data being both OAR and MAR (Holman & Glas, 2005). Relating this to the missingness means that the probability of a missing response depends neither on the observed variable nor on any unobserved variables. If the probability of a missing response depends on the observed variable—including other administered covariates—but is not influenced by an unobserved variable or process, the missingness is labeled MAR. It is not completely at random, since the data are not observed at random, meaning that they depend on the variables that are under observation. In case the probability to observe a missing

response differs for the same value of the observed variable and this mechanism cannot be explained by either the observed or any of the recorded covariates, the observed missingness in the data is called MNAR. These unobserved variables might influence the occurrence of missing data, and thus the inferences drawn from the observed data matrix might be biased. As an example, the tendency to report one's body weight might depend on how comfortable a person feels about their appearance. Now contemplate a study on the relationship between income and body weight. If the researcher disregards the fact that the data are not missing at random and only information from people who feel comfortable with their body weight are recorded, the inferences drawn from this analysis are bound to be biased. One possible conclusion could be that the income does not depend on body weight, when in fact a complete data matrix would reflect a relationship between the two. If, however, the research design includes a question about the contentedness of one's appearance and these values are considered in the data analysis, the MAR assumption might hold.

In summary, if the probability for the observed missing data matrix equals the probability for the missing data matrix as a function of the observed variable, missing variables, covariates, and the latent trait, the unobserved data is MCAR. If the missingness in the observations does depend on the observed variable and/or other covariates, the missing values are MAR, and in case the MAR assumption does not hold, they are MNAR (see Holman & Glas, 2005; Mislevy & Wu, 1988, 1996; Schafer & Graham, 2002).

In terms of ignorability, missing values can be ignored in the scaling when they are (a) either MCAR or MAR, and when (b) the parameter of the missing data process—that is, the distribution of answered and unanswered items—is distinct from the parameter of the data process—that is, the distribution of correct and incorrect responses (Mislevy & Wu, 1988, 1996). Distinctness refers to independence between the parameter space of the missing data process and the parameter space of the observed data process (Mislevy & Wu, 1988, 1996).

Missing values are considered ignorable when (a) and (b) are met; they are considered nonignorable when either (a) or (b) are violated (Rubin, 1976; Mislevy & Wu, 1988, 1996).

Note that in principle, not the data but the variables included in the analysis determine whether the results are influenced by the missing data (Schafer & Graham, 2002). If the missing data mechanism is considered and a model for this mechanism is established, correct inferences can be drawn from the data (Rubin, 1976).

1.2.2 Ignorability of missing data in large-scale assessments

In large-scale assessments, not-administered items can be considered MCAR, since the missing data mechanism neither depends on observed nor unobserved variables (Mislevy & Wu, 1988, 1996). The missingness in the data is completely determined by the researcher and the random assignment of the alternate test forms to the examinees. Therefore, missing values due to not-administered items can be ignored (Mislevy & Wu, 1988, 1996).

Not-reached items typically occur due to a time limit of the test. Most large-scale tests do not intend to measure speediness, therefore aiming to distinguish between the actual ability parameter and the speed parameter (Lord, 1974). Basically, the amount of not-reached items can be considered the speed parameter, so simply ignoring the not-reached items would solve the problem of separating the two (Lord, 1974). However, the speed parameter is usually not distinct from the ability parameter, meaning that the expected level of speediness is different for different ability levels (Koretz, Lewis, Skewes-Cox, & Burstein, 1993; Pohl, et al., 2014). Empirical studies have shown that the amount of not-reached items often depends on the ability of the person (Glas & Pimentel, 2008; Koretz et al., 1993; Pohl, et al., 2014). Thus, contrary to the assumptions outlined in the previous chapter, the amount of not-reached values does hold information about the ability of the person, and is therefore nonignorable. Ignoring missing values and not taking the relationship between speediness and ability into account can lead to incorrect inferences on ability (Mislevy & Wu, 1988, 1996). Note that this

nonignorability hardly received attention in practice, and most large-scale assessments either ignore not-reached items or treat them as incorrect.

In terms of omitted items, the missing data mechanism is even more difficult to specify. Numerous parameters constitute to the omission process, including characteristics of the test, characteristics on the item-level, and characteristics on the person-level (O'Muircheartaigh & Moustaki, 1999). Influences on the test-level comprise the scoring method, for example, whether a penalty for incorrect scores exists, and also whether examinees are aware of how omissions are treated (Sabers & Feldt, 1968). Furthermore, the type of test—whether high-stakes or low stakes—as well as incentives for testing might play a role in why and how many items are omitted (Berlin et al., 1992). On the item-level, item difficulty is one of the main influencing factors of the amount of omissions. Many studies found that more difficult items are skipped more frequently (Koretz et al., 1993; Pohl et al., 2012; Rose et al., 2010; Zhang, 2013). Another relevant characteristic on the item-level is the response format. In the 1990 NAEP study, open-ended tasks were skipped more often (Koretz et al., 1993), which might be partly due to them being amongst the most difficult, but also because guessing on these items is more challenging than guessing on a multiple-choice item. Lastly, individual person characteristics greatly determine the missing data mechanism. People differ in their general tendency to omit items, and thus have different thresholds for omitting an item (Jakwerth, Stancavage, & Reed, 1999; Zhang, 2013). Also, examinees employ different test-taking strategies; they are differently motivated and may differ in their confidence about their answers (Jakwerth et al., 1999). Furthermore, many studies have demonstrated a dependency between omissions and ability (Pohl et al., 2014; Rose et al., 2010; Stocking, Eignor, & Cook, 1988; Zhang, 2013). Overall, test, item, and person specific factors constitute the omission process. In terms of ignorability of omissions, MAR would hold if the probability of a missingness pattern were constant across all ability levels (Mislevy & Wu, 1996). As previously stated, this is hardly the case in empirical data, since people with

higher abilities are more prone to respond to items. Therefore, omitted items are also nonignorable (Holman & Glas, 2005; Mislevy & Wu, 1996). Contrary to the assumptions outlined in the previous chapter, however, lack of knowledge only comprises part of the reason for an omission. Treating all omitted items as incorrect responses reduces the omission process to one single underlying cause, when in reality several other factors play a role in the missing data mechanism for omitted items. Also, the assumption that an omitted item has been reached and sufficiently pondered by the examinee does not necessary hold. Jakwerth et al. conducted a qualitative investigation on reasons for missing values, discovering that some examinees skip items before actually putting time and effort into solving the item. This is an additional indicator that not all omitted items are items that were simply too difficult for the examinee.

All in all, the mechanism underlying the missing data and the extent to which the missing data actually depend on ability needs to be considered in order to make accurate inferences about person ability estimates.

1.3 Model-Based Approaches for Nonignorable Nonresponse

> "Not surprisingly, modeling this nonignorable nonresponse is difficult"
>
> (Mislevy & Wu, 1988, p. 62)

As previously established, not-reached and omitted items pose a threat to accurate ability measurement, since their occurrence is systematic and might thus distort parameter estimation (Lord, 1974; Mislevy & Wu, 1996; Rubin, 1976). The commonly employed missing data approaches do not account for missing values that are MNAR. In the last several decades, researchers developed model-based approaches in order to account for this nonignorable nonresponse. The following section introduces these approaches, illustrating their basic concept and outlining different types (1.3.1). It subsequently discusses the performance of model-based approaches, especially with regard to large-scale assessment studies (1.3.2).

1.3.1 Introduction to model-based approaches

The term *model-based approach* applies to certain types of models. The novelty of these models is the idea of obtaining a more accurate measure of ability by including information of the missing data process in the measurement model (i.e., the scaling model) for ability. The method is similar to using collateral information, such as covariates, in order to reduce estimation error. A model needs to be established in order to model the missing data process (Rubin, 1976). This model typically comprises a latent or manifest variable, the *missing propensity*, which represents a person's general tendency to respond to an item (Glas & Pimentel, 2008; Glynn et al., 1986; Heckman, 1976; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Rose, et al., 2010). Basically, the missing propensity serves as a covariate that is accounted for when including it in the measurement model. The ability score can thus be considered an augmented score. The amount of augmentation taking place depends on the extent of the correlation between ability and missing propensity. If they are unrelated, the ability scores will not be affected by the missing propensity.

The models can be classified according to (a) the type of missing items they account for, and (b) whether they use a manifest or a latent approach to model the missing propensity. With regard to (a), the two types of missing responses typically considered are omitted responses and not-reached responses (see, e.g., Glas & Pimentel, 2008; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Rose, et al., 2010). In terms of (b), the manifest approach includes the missing propensity in form of a manifest variable that consists of the aggregated number of missing values of a person (Rose et al., 2010). This variable can then be included in a latent regression model (see Figure 1).

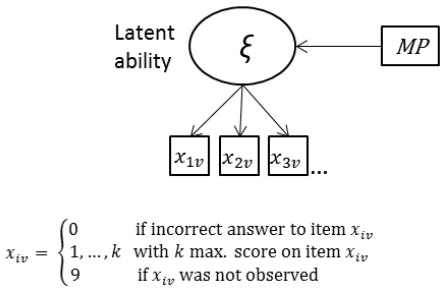

Figure 1. Manifest model-based approach. MP = missing propensity. *i* indexes the items from *i* = 1, …, *I*, and *v* indexes the persons from *v* = 1, …, *V*.

In the latent approach, the missing propensity is modeled as a latent variable, with manifest missing data indicators d_{iv}. The missing data indicators are commonly defined as

$$d_{iv} = \begin{cases} 0 & \text{if } x_{iv} \text{ was not observed} \\ 1 & \text{if } x_{iv} \text{ was observed.} \end{cases} \quad (1)$$

The missing-data indicator matrix, which contains the missing data indicators, is made up of the same number of persons and the same number of items as the matrix of the observed data. The models differ in how they incorporate the latent missing propensity in the measurement model. Some prominent approaches are selection models (Heckman, 1976) and pattern mixture models (Glynn et al., 1986; Rubin, 1987), though they are rarely applied in practice due to challenges in parameter specification and identification (O'Muircheartaigh & Moustaki, 1999). O'Muircheartaigh and Moustaki proposed a latent within-item-multidimensional IRT (W-MIRT) model, where the missing data indicators load on both the ability and the missing propensity (see Figure 2a). Holman and Glas (2005) extended this model, proposing various W-MIRT as well as between-item-multidimensional IRT (B-MIRT) models. In B-MIRT models, the missing data indicators solely load on the missing propensity (see Figure 2b).

a. b.

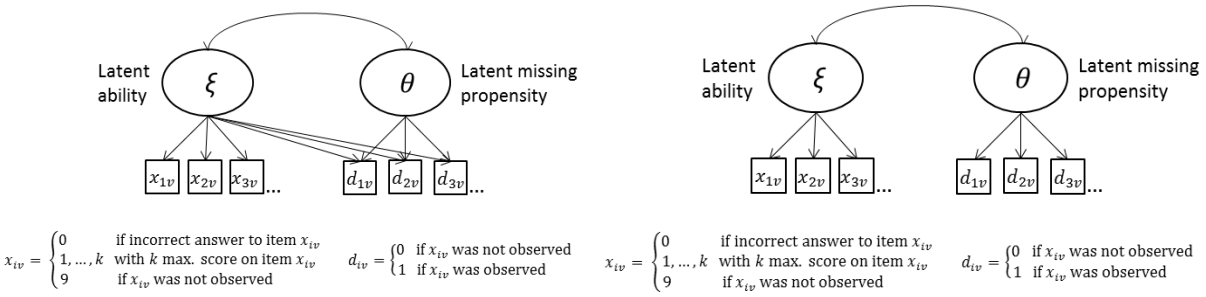

Figure 2. Latent within-item-multidimensional IRT (W-MIRT) model (a) and latent between-item-multidimensional IRT (B-MIRT) model (b) to account for nonignorable nonresponse.

In order to obtain the parameters for these multidimensional IRT models, marginal maximum likelihood (MML) is applied. This method requires an assumption for the joint distribution of ability and missing propensity, which is typically assumed to be bivariate normal (O'Muircheartaigh & Moustaki, 1999).

Due to the simple structure in B-MIRT models, the interpretation of the missing propensity is more straightforward than in W-MIRT models where multiple variables load on the missing data indicators (Rose, 2014; Rose et al., 2010). They are thus recommended for application in practice (Rose, 2014; Rose et al., 2010). Another aspect that needs consideration when choosing between different model-based approaches is how to model the missing propensity. The probability of examinee *v* to respond to item *i* can either be modeled using the 2PL model (Birnbaum, 1968) or the Rasch model (Rasch, 1960). In the former, the item loadings—or, in IRT terms, slope/discrimination parameters—can vary freely, in the latter they are fixed at 1. In most competence tests, the amount of missing values is rather limited, and therefore the variance of the missing data indicators is usually small (Holman & Glas, 2005). This might result in convergence problems for estimating all parameters of the 2PL model (Pohl et al., 2014). Fixing the slopes at 1 is therefore more convenient. The Rasch model is thus often preferred when modeling the latent missing propensity.

Glas and Pimentel (2008) developed a latent approach to specifically account for not-reached items. They use a sequential model with linear restrictions on the item difficulty parameters in order to adequately map the features of the propensity to not reach items. Further developments include (a) models that simultaneously take omitted and not-reached items into account (Pohl et al., 2014; Rose, 2013), (b) models that include covariates (Glas, Pimentel, & Lamers, in press; Moustaki & Knott, 2000; Rose, 2013), and (c) approaches that model the missing propensity two-dimensionally (Glas et al., in press; Rose, 2013). These three additional features of the models were considered since (a) omitted and not-reached items both need to be accounted for, but the respective missing processes might be distinct from each other, which therefore requires separate models (Pohl et al., 2014; Rose, 2013), (b) the missing process in competence tests relates to other person characteristics (Jakwerth et al., 1999; Matters & Burnett, 2003; Zhang, 2013), which may need to be included in order to appropriately model the missing process, and (c) the missing process is possibly multidimensional and should be modeled accordingly (Rose, 2013).

In sum, several model-based approaches exist, which vary in how the missing mechanism is modeled. The approach may prove to be a superior alternative to the commonly employed approaches for estimating competence scores, since the models allow accounting for nonignorable nonresponse. Some of the model-based approaches have been scrutinized in light of their applicability to large-scale assessments, which will be discussed next.

1.3.2 Performance of model-based approaches

Whether model-based approaches serve as a superior alternative to handle missing data and whether they are applicable to large-scale assessments has been investigated in a number of simulation studies and real data applications (see, e.g., Holman & Glas, 2005; Pohl et al., 2014; Rose, 2013; Rose et al., 2010). In simulation studies, the first step is to generate complete data sets using fixed item and person parameters. In the second step, missing values are induced according to a certain missing data mechanism in order to obtain data sets with

missing data. Several studies vary the extent to which the missing data mechanism depends on the ability. Note that the size of the correlation between missing propensity and ability is considered to index the amount of nonignorability in the data (Holman & Glas, 2005; Rose et al., 2010). If missing propensity and ability are unrelated, the conditions of ignorability are met and the missing values can be ignored (Holman & Glas, 2005).[4] Lastly, the data sets with missing data are analyzed using different missing data approaches, and the estimated item and person parameters are compared to the true parameters. When analyzing real data, the item and person parameter estimates from different missing data approaches are simply compared. Also, reliabilities of person parameter estimates and model-fit indices are typically used to evaluate the performance of different models. Note that under this approach, the true values of the item and person parameters are unknown.

Holman and Glas (2005) applied their B-MIRT and W-MIRT approaches to simulated data and data from a medical disability study. Glas and Pimentel (2008) investigated their model for not-reached items in a simulation study, and used data from a speeded intelligence test to illustrate the applicability of their method. Rose et al. (2010) considered the manifest approach as well as B-MIRT and W-MIRT models using simulated data and data from the PISA 2006 study. Pohl et al. (2014) applied the approaches by Holman and Glas, Glas and Pimentel, as well as their own models, which simultaneously account for not-reached and omitted items, to NEPS competence data, validating their results in a simulation study. Despite investigating the performance of model-based approaches, most of the mentioned studies also applied the common missing data approaches of ignoring missing values and treating them as incorrect. The studies unanimously conclude that model-based approaches

[4]Note that in real data, the assumption of MAR cannot be proven (Schafer & Graham, 2002). Therefore, even if ability and missing propensity are unrelated, the conditions for ignorability might not be met, since the missing data possibly depend on another unknown mechanism beyond the researcher's knowledge and control. Follow up studies can help identify hidden reasons for nonresponse (Glynn, Laird, & Rubin, 1993).

perform well both in simulated and in real data settings.[5] In the simulations, item and person parameter estimates hardly deviate from the true parameters. In the real data settings, the models show good model fit and high reliability of the ability estimates. With regard to the common missing data approaches, the approach of ignoring missing values leads to quite analog item and person parameter estimates as the model-based approaches. The approach is slightly inferior to model-based approaches in terms of model fit and reliability. Treating missing values as incorrect responses drastically biases item and person parameter estimates, and leads to overestimated reliabilities.

Overall, including the missing propensity in the measurement model for ability seems a very promising approach to account for nonignorable nonresponse in large-scale assessments. So far, the approaches are not implemented as a scaling method for competencies in large-scale assessments. Several aspects need further investigation. These include (a) testing whether the assumptions of model-based approaches actually hold in empirical settings, (b) determining which model-based approach best describes the missing mechanism in competence tests, and (c) evaluating the feasibility as well as the necessity of model-based approaches compared to the common, computationally less extensive approaches.

1.4 Research Topics

"The conclusions reached seem very plausible
in view of the theoretical results proved in the Appendix.
Further empirical confirmation would be desirable also"
(Lord, 1974, p. 285)

The overall aim of this dissertation was to investigate the properties of the missing data process and to find a scaling model which adequately takes missing responses into

[5]This sentence as well as the following sentences within the paragraph refers to studies from Holman and Glas (2005), Pohl et al. (2014), Rose (2013), and Rose et al. (2010).

account. Three comprehensive studies were conducted, which centered around these research foci. The following section discusses the specific research questions investigated in each study. The first study focused on the assumptions made in model-based approaches, whether they hold in empirical settings, and how violations to these assumptions influence person parameter estimates (1.4.1). The second study comprises substantial research on the missing-data mechanism in competence tests (1.4.2). The third study aims at answering the question of an adequate missing data approach for large-scale assessments (1.4.3). The last section connects the three studies, embedding them in the theoretical framework and connecting them to the overall research aim of the dissertation (1.4.4).

1.4.1 Assumptions of model-based approaches

As stated earlier, the aim of model-based approaches is to include the missing data mechanism in the measurement model for ability. By using the information of the nonresponse, nonignorable missing values are taken into account. This approach requires a model for the missing propensity. Establishing an adequate model is a necessary prerequisite for applying the model-based approaches to large-scale assessment studies. Only models making adequate assumptions about the missing data mechanism can reliably take nonignorable nonresponse into account.[6]

The first research paper entitled “Taking the missing propensity into account when estimating competence scores – Evaluation of IRT models for non-ignorable omissions” deals with latent models for omitted items.[7] Two key assumptions of the between-item-

[6]For more detailed information and references regarding this paragraph, see 1.3.1.

[7]Note that in his dissertation, Rose (2013) argues to only apply the manifest model-based approach to not-reached missing values, since the occurrence of a not-reached item depends on whether or not the previous item has been reached or not. Therefore, the missing data indicators for not-reached items violate the assumption of local stochastic independence in MIRT models, and the missing propensity to not reach items should not be modeled using the latent model-based approach. Accordingly, when investigating the latent model-based

multidimensional IRT model to account for nonignorable missing values (Holman & Glas, 2005) and whether they hold in empirical settings were investigated: (1) the assumption of unidimensionality and (2) the assumption that the latent variables ability and missing propensity are bivariate normally distributed.

The motivation behind the first research question was that up to that point, studies applying the latent model-based approach modeled the missing propensity unidimensionally (Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010). This underlies the implicit assumption that the latent missing propensity is, in fact, unidimensional. Rose (2013) states that modeling the missing propensity unidimensionally although the true underlying missing data process is actually multidimensional can result in biased parameter estimates. The model-based approach might thus fail at accurately accounting for nonignorable nonresponse. Models for a multidimensional missing propensity have been proposed, but not applied to empirical data (Glas et al., in press; Rose, 2013). Thus far, hardly any studies investigated whether the missing propensity is actually unidimensional and can be modeled accordingly. Pohl et al. (2014) explored several item-fit indices, which indicated that the missing data indicators for omitted items fit the Rasch model (Rasch, 1960) well. Besides these investigations, no research was conducted on whether examinees' missing propensity in competence tests of large-scale assessments can be modeled as a unidimensional latent variable.

Besides the assumption of unidimensionality of the missing propensity, another implicit assumption when estimating the parameters for the latent multidimensional model-based approach concerns the joint distribution of the two latent variables ability and missing propensity. As stated earlier, the MML method is typically applied under the assumption of a

---

approach, emphasis is based on the latent propensity to omit items. This also concerns the studies conducted within the scope of the present dissertation.

multivariate normal distribution (see 1.3.1). However, the amount of omissions per person is typically not normally distributed (e.g., Pohl et al., 2012). Assuming normality in the MML estimation procedure despite an underlying skewed distribution can result in biased item and person parameter estimates (Molenaar, 2007; Stone, 1992; Zwinderman & van den Wollenberg, 1990). How and whether a violation to the bivariate normal distribution assumption affects the performance of model-based approaches has not been investigated thus far.

In our study we proposed models with more lenient distribution assumptions and compared them with the model by Holman and Glas (2005). After establishing the model that makes appropriate assumptions in the context of real data, we examined whether the ability parameters from this model greatly deviated from ability parameters from a model that simply ignores missing values. This served as an indicator of whether the model that ignores missing values is robust against violations of ingnorability and produces accurate ability estimates. The overall purpose of the study was to evaluate the model-based approach and to establish its feasibility to real data. If the assumptions are violated, the models might need to be altered or extended in order to ensure applicability.

1.4.2 Features of the missing propensity in competence tests

Large-scale assessments typically test examinees in several competence domains. These domains are often scaled simultaneously, resulting in a multidimensional IRT model. A propensity to not reach items and a propensity to omit items exists in each domain. Including the missing propensities to omit items as latent variables in the model would drastically increase the number of dimensions, making the model computationally extensive (see, e.g., Rose, 2013). A favorable solution would be to model the omission propensity only once. Beside this major advantage with regard to the scaling of the omission propensity, the substantive meaning of a stable propensity to omit and not reach items is of interest. If the missing propensity can, in fact, be considered a person's general tendency to omit or not reach

items, respectively, it should persist irrespective of test content. The person-specificity of the missing propensity might further be reflected in a cross-domain stable relationship to other person inherent characteristics.

The second paper entitled “Investigating mechanisms for missing responses in competence tests” focusses on features of the missing propensity itself. It investigates (1) to what extent a person’s missing propensity varies across different competence domains, and (2) whether the missing propensity is related to other stable person characteristics such as demographic variables and competencies.

The study closes the gap to previous research, where the application of model-based approaches mainly focused on the missing propensity in a single competence domain, for example, the ability in reading or mathematics. Hardly any information exists on persons’ response rates across different domains.

The influences of person characteristics on the missing mechanism have been investigated to some extent. Several studies reported dependencies between the amount of missing values and demographic variables as well as motivation (Jakwerth et al., 1999; Koretz et al., 1993; Matters & Burnett, 2003; Zhang, 2013). However, these studies were rather limited in the number of possible explaining variables and considered only a single competence domain and only one age cohort.

We examined both research questions in an extensive study, analyzing several competence domains, several age cohorts, and numerous possible explaining variables. The study served to better understand the mechanisms that lead to omitted and not-reached items in actual competence tests. Moreover, the aim was to find general, stable patterns of the missing mechanism in competence tests. This information is necessary in order to accurately take peoples’ missing propensities into account. If the missing propensity is rather stable and relates to other stable person characteristics, it would not have to be investigated for each competence domain separately, and it might suffice to only include it in the scaling model

once. Furthermore, a clearer understanding of the missing mechanism in real data might be useful for simulation studies investigating missing data. The advantage of simulation studies over real data is that in simulation studies, the true parameters are known. Simulation studies may thus aid to determine adequate missing data approaches, since the parameters estimated from different missing data approaches can be compared with the true parameters. For these studies, however, the missing values need to be induced according to a certain mechanism. Closely mapping this mechanism to the actual occurrence of missing values in real data will allow for more generalizable and valid inferences from simulation studies.

1.4.3 Necessity of model-based approaches in large-scale assessments

After establishing how to accurately include the missing data mechanism in the measurement model for ability, the question remains when these models should be implemented in empirical settings. How meaningful is the gain in accuracy compared to the common approaches? How do possible violations to assumptions affect parameter estimates? Further, the complexity and computational costs of the model-based approaches need consideration. Previous studies found no substantial deviations between ability estimates from models including the missing propensity and models simply ignoring the missing data (Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010). Whether and to what extent nonignorable missing values influence ability estimates depends on features of the missing data mechanism in the respective data. One influencing factor is the amount of nonignorable missing values in the data (Rose et al., 2010). Also, the size of the correlation between the missing values and ability as well as between the missing values and other observed variables affect the accuracy of ability estimates (Glas et al., in press; Holman & Glas, 2005; Rose et al., 2010). In order to establish under which missing data conditions nonignorable missing values threaten the reliability of competence estimation, all relevant influencing factors need to be considered corporately.

In the third research paper entitled “Performance of missing data approaches in retrieving group-level parameters” we investigate how well different missing data approaches perform in retrieving a regression coefficient when the missing propensity relates to the explaining variable (i.e., the predictor) and ability.

The study was motivated by the fact that thus far, individual ability scores, reliability estimates of these scores, and model comparisons served as indicators to evaluate different missing data approaches (see 1.3.2). Group-level scores have not been investigated, even though comparisons between groups of people or relationships between ability and other variables are typically of interest when analyzing competence test data of low-stakes assessments. An important aspect when estimating relationships on the group-level, for example, between ability and a third variable, is that this third, explaining variable might correlate with the missing data mechanism. These considerations raise the question of an adequate missing data treatment when the missing mechanism depends on ability and the explaining variable. This issue has not been considered so far, especially not with respect to the parameter estimate on the group-level.

Our objective was realized by conducting a simulation study in which we varied the amount of missing data as well as the mechanism inducing the missing data. We then analyzed the data sets containing missing data with three missing data approaches: (a) treating missing data as incorrect, (b) ignoring missing data, and (c) applying the B-MIRT model proposed by Holman and Glas (2005). We subsequently compared the true regression coefficient with the estimated coefficients from the different approaches in order to evaluate which approach is superior. This enables us to evaluate the necessity of the model-based approach in large-scale assessments. It allows establishing under what missing data conditions the bias on the group-level becomes large, and under which conditions simpler missing data approaches suffice to obtain accurate parameter estimates.

1.4.4 Summary and holistic view of the three research foci

Reverting back to the beginning of the chapter, I established that the goal of large-scale assessments is to provide data that allow answering substantive research questions regarding educational systems (see 1.1.1). Generally, the data sets from the conducted studies contain missing values. Especially missing values due to not-reached items and due to omitted items threaten the validity and reliability of the test, as they usually correlate with the unobserved responses (see 1.2.2). They are thus MNAR, and ignoring them can lead to biased parameter estimates (see 1.2.1). So far, most large-scale assessment studies deal with not-reached and omitted items by ignoring them or treating them as incorrect (see 1.1.2). Since the assumptions underlying these missing data approaches are irreconcilable with empirical evidence of the actual missing data process, neither of the approaches can be considered state-of-the-art. These considerations point out the major gap between research and practice. The development of newer model-based approaches might bridge this gap, although they have hardly been investigated in light of their applicability to large-scale assessment data. A number of aspects need further consideration, which are investigated in the present dissertation.

First, the properties of the missing data process in actual competence test data had to be examined. This was necessary in order to validate the applicability of the model-based approach to large-scale assessment data, and to ensure that the missing propensity reliably captured the missing data mechanism. The first study focused more on technical matters that were rather specific to the assumptions of the multidimensional IRT model of the latent model-based approach. It answered the question of whether the assumptions of the model-based approaches for the missing data mechanism hold in empirical settings. The second study also dealt with the properties of the missing data mechanism, but centered predominantly on the substance and the influencing factors of the missing data mechanism.

This study helped identifying relevant features of the missing mechanism in large-scale assessments.

Second, an adequate scaling model for large-scale assessments needed to be established. In light of this aim, the different missing data approaches, including model-based approaches and the more common approaches, were evaluated. The first two studies were a necessary prerequisite for this evaluation. They developed a model that adequately takes nonignorable nonresponses in competence tests into account, thus ensuring the reliability and validity of the model-based approach. In the third study, bias of parameter estimates on the group-level served as the criterion for evaluating the different approaches. Selecting this criterion was based on the fact that analyses using data from large-scale assessments are typically conducted on the group-level. Different possible missing data scenarios were considered in order to make general statements about the important research question: Under what missing data conditions do which approaches lead to sufficiently accurate estimates? The study also points out consequences and possible false conclusions from data analyses that handle missing data inadequately. All in all, the third research paper answers questions relevant for the scaling and analyzing of competence tests in large-scale assessments.

In sum, the aim of the dissertation is to aid in bridging the gap between theory (i.e., the theoretically derived conditions on when missing values can be ignored and the postulated models to incorporate the missing data mechanism) and practice (i.e., the occurrence and handling of missing values in the scaling of competence tests). Both aspects are considered in order to provide a comprehensive view on the matter, which might allow deriving conclusions about how missing values should best be dealt with in competence tests of large-scale assessment studies.

Chapter 2

Study 1: Taking the missing propensity into account when estimating competence scores—Evaluation of IRT models for non-ignorable omissions

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores – Evaluation of IRT models for non-ignorable omissions. *Educational and Psychological Measurement, 75(5),* 850–874.

http://dx.doi.org/10.1177/0013164414561785

Educational and Psychological Measurement
2015, Vol. 75(5) 850–874

DOI: 10.1177/0013164414561785
epm.sagepub.com
SAGE

2 Taking the missing propensity into account when estimating competence scores—Evaluation of IRT models for non-ignorable omissions

Carmen Köhler[1], Steffi Pohl[2], and Claus H. Carstensen[1]

[1]Otto-Friedrich-University Bamberg, Germany

[2]Free University Berlin, Germany

Author Note

Carmen Köhler, Otto-Friedrich-University Bamberg, Germany; Steffi Pohl, Free University Berlin, Germany; Claus H. Carstensen, Otto-Friedrich-University Bamberg, Germany.

This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 4–9th Grade (School and Vocational Training - Education Pathways of Students in 9th Grade and Higher), doi:10.5157/NEPS:SC4:1.0.0. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

Correspondence concerning this article should be addressed to Carmen Köhler, Otto-Friedrich-University Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany. Phone: +49-951-863-3406. E-Mail: carmen.koehler@uni-bamberg.de

Abstract

When competence tests are administered, subjects frequently omit items. These missing responses pose a threat to correctly estimating the proficiency level. Newer model-based approaches aim to take non-ignorable missing data processes into account by incorporating a latent missing propensity into the measurement model. Two assumptions are typically made when using these models: (1) The missing propensity is unidimensional, and (2) the missing propensity and the ability are bivariate normally distributed. These assumptions may, however, be violated in real data sets, and could, thus, pose a threat to the validity of this approach. The present study focuses on modeling competencies in various domains, using data from a school sample (N = 15,396) and an adult sample (N = 7,256) from the National Educational Panel Study. Our interest was to investigate whether violations of unidimensionality and the normal distribution assumption severely affect the performance of the model-based approach in terms of differences in ability estimates. We propose a model with a competence dimension, a unidimensional missing propensity and a distributional assumption more flexible than a multivariate normal. Using this model for ability estimation results in different ability estimates compared to a model ignoring missing responses. Implications for ability estimation in large-scale assessments are discussed.

Keywords: missing data, non-normal distribution, Item Response Theory, scaling competencies, large-scale assessment

Theoretical Background

In the late 1950s, the interest in comparing students´ skills on the national as well as the international level led to the onset of the large-scale assessment era (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962). In order to enable educational monitoring, data on student knowledge is systematically collected via competence tests. These large-scale assessment studies allow investigating complex research questions in the educational field concerning educational processes, competence development, and educational decisions. *Item Response Theory* (IRT) has manifested itself as the psychometric basis for scaling the competencies in large-scale assessments (von Davier, Gonzalez, Kirsch, & Yamamoto, 2013). In IRT, the answers to questions in a competence test serve as indicators of the participant´s latent proficiency, allowing the researcher to draw inferences from the manifest response behavior on the underlying, unobservable trait. The concept of measuring a construct becomes more complicated when some of the manifest indicators are missing due to examinees skipping parts of the test. Incomplete data impedes drawing correct inferences on the trait to be measured, since some of the required information remains missing, and the missing values may be non-ignorable (Mislevy & Wu, 1996).

Of course, the impact missing values have on the scaling of proficiencies depends on the amount of their occurrence. Large-scale assessment studies distinguish between different types of missing values, which vary in frequency. Some items are usually *missing by design*, since not all test items are administered to each subject. When a participant gives an answer not listed among the options, the answer is coded *invalid*. *Not reached* items are questions the participant did not answer due to time limits. Missing items which the examinee chose to skip are labelled *omitted*. Although large-scale studies aim at giving the participants ample time for the completion of the test and no penalty for guessing results, examinees still show a remarkable amount of missing data. Whereas invalid answers hardly occur, the amount of omitted and not reached items is more striking. For example, in the Programme for

International Student Assessment (PISA) 2000 study, the average number of omitted competence items of the second testing session exceeded 5% in six of the participating countries (Adams & Wu, 2002). These findings were similar regarding not reached items. Data from the 1990 NAEP study in grade twelve shows that for 9% of the mathematics items, omission rates exceeded 10%; these numbers were comparably higher for not-reached items (Koretz, Lewis, Skewes-Cox, & Burstein, 1993).

So far, researches have not reached a consensus on how to ideally manage unobserved values in IRT models, and various large-scale studies employ different approaches on treating missing data. In PISA (Adams & Wu, 2002) and the Third International Mathematics and Science Study (TIMSS; Martin, Gregory, & Stemler, 2000), a two-stage procedure is employed, where missing values are ignored in item calibration, but treated as incorrect when estimating person ability parameters. Other studies use different strategies for different types of missing responses. In NAEP (Johnson & Allen, 1992), for example, not-reached items are ignored, while omitted items are scored as fractionally correct, using the reciprocal of the number of response options of the multiple-choice item as the response value. In the National Educational Panel Study (NEPS; Pohl & Carstensen, 2012), all missing responses are ignored in the scaling, meaning those items are considered as having not been administered to the participant. Another possibility of dealing with unobserved items—though not commonly applied to large-scale assessments—involves imputing the missing values via two-way imputation (Bernaards & Sijtsma, 2000), response-function imputation (Sijtsma & van der Ark, 2003), conditional mean imputation (Schafer & Schenker, 2000), the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), or multiple imputation (MI; Rubin, 1987).

Many studies have investigated the performance of the aforementioned methods, illustrating their strengths and limitations. In 1974, Lord already argued that treating omitted items as *wrong* leads to biased parameter estimates. Several simulation studies support this

statement, while also concluding that substituting an incorrect value for a missing answer creates more bias than simply ignoring omissions (see, e.g., de Ayala, Plake, & Impara, 2001; Hohensinnn & Kubinger, 2011). Results from Finch (2008), who compared several imputation techniques as well as the traditional approaches, indicate that the least ideal method is to treat omits as *wrong*, while none of the other methods differed substantially in their performance. In a simulation study conducted by Culbertson (2011), ignoring missing responses or treating them as fractionally correct outperformed the EM algorithm, the MI approach, and scoring omits as wrong.

All these approaches can handle missing responses only if (a) the missing responses are either *missing completely at random* (MCAR) or *missing at random* (MAR), and if (b) the parameter vector of the probability density function of the missing-data matrix is distinct from the parameter vector of the probability density function of the complete data matrix (Rubin, 1976). With regard to competence test data, both conditions are usually violated, which may result in biased ability estimates (see, e.g., Mislevy & Wu, 1988, 1996).

Since the different types of missing values that generally occur in large-scale assessments result from different missing processes, the ignorability of the missing responses need to be investigated separately for each missing type. When items are missing due to the design, the researcher can control for the process which led to the missing data. This is possible because the process causing the missing data is known. MAR or even MCAR, as well as distinctness hold for this type of missing, and the missing responses can therefore be ignored (Mislevy & Wu, 1996). For not reached and omitted items, the MAR and distinctness assumption are typically violated. Many studies found that these types of missing responses relate to the ability of the person (e.g., Glas & Pimentel, 2008; Koretz et al., 1993; Rose, von Davier, & Xu, 2010; Stocking, Eignor, & Cook, 1988). The probability for omitting or not reaching an item depends not only on the difficulty of the item, but additionally on the

unobserved latent trait, ξ. Thus, both MAR and distinctness are violated. The process leading to the missing values is therefore not ignorable and needs to be accounted for.

The current paper focusses on missing responses which are due to omissions, and draws on a model-based approach developed by O`Muircheartaigh and Moustaki (1999) and extended by Holman and Glas (2005). This particular approach tries to take non-ignorable omissions into account by jointly modeling the distribution of the ability and the missing propensity (Holman & Glas, 2005; O`Muircheartaigh & Moustaki, 1999). Let v index the person, for $v = 1,\ldots,V$, and i index the test item, for $i = 1,\ldots,I$. The ability is modeled on the basis of the matrix $\boldsymbol{X}$, which contains the observed values x_{iv}. O`Muircheartaigh and Moustaki (1999) define the second dimension, the *response propensity*, θ, as a latent variable "which represents a general tendency to respond, varying across individuals" (p. 179). This latent variable is modeled on the basis of the matrix $\boldsymbol{D}$, which consists of the missing data indicators d_{iv}, and is built up of the same number of both i and v as the matrix $\boldsymbol{X}$. The missing data indicators can be defined as

$$d_{iv} = \begin{cases} 0 & \text{if } x_{iv} \text{ was not observed} \\ 1 & \text{if } x_{iv} \text{ was observed,} \end{cases} \quad (1)$$

so that for each missing value x_{iv} in $\boldsymbol{X}$, $d_{iv} = 0$. Note that higher missing propensity values indicate less missing responses. In order to make inferences on examinee proficiency while accounting for non-ignorable non-response, a measurement model on the probability of observing a response and a model on the probability of giving a correct answer are combined to form a multidimensional IRT (MIRT) model. The model-based approach allows incorporating both ability and missing propensity, as well as further covariates into the same multidimensional measurement model, estimating the parameters of interest in a one-stage procedure. They are in turn very flexible, and also combine all information simultaneously

(Moustaki & Knott, 2000). Holman and Glas (2005) propose various MIRT models accounting for omitted responses, including within-item-MIRT (W-MIRT) models along with between-item-MIRT (B-MIRT; see Figure 1) models. In W-MIRT models—which encompass the model proposed by O`Muircheartaigh and Moustaki (1999)—the missing data indicators load on both ξ and θ, whereas in B-MIRT models they solely load on θ. Thus, the difference between the two models resides in the fact that in W-MIRT models the probability of observing a response is modeled as a function of ξ, θ, and δ_i—with δ_i denoting the difficulty of giving an answer to the item *i*—whereas in B-MIRT models the probability of observing a response is modeled as a function of only θ and δ_i. Rose et al. (2010) discuss the equivalence of B-MIRT and W-MIRT Rasch models, but additionally demonstrate that the latent variable θ in B-MIRT models has a different meaning in W-MIRT models and cannot truly be considered a response propensity in the latter. The authors therefore recommend applying the B-MIRT model to account for non-ignorable omissions.

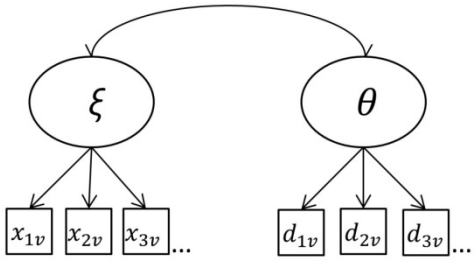

Figure 1. Between-item-multidimensional-IRT model to account for non-ignorable omissions.

The marginal maximum likelihood (MML) of the B-MIRT model is given by

$$L = \prod_{v=1}^{V} \prod_{i=1}^{I} p(x_{iv}|\xi_v, \beta_i) p(d_{iv}|\theta_v, \delta_i) g(\xi_v, \theta_v|\boldsymbol{\phi}), \quad (2)$$

where $p(x_{iv}|\xi_v, \beta_i)$ represents the probability that person *v* gives a correct response to item *i* as a function of person ability ξ_v and item difficulty β_i; $p(d_{iv}|\theta_v, \delta_i)$ represents the

probability of observing an answer from person *v* on item *i* as a function of the persons missing propensity θ_v and the difficulty of giving an answer to item *i*; $\boldsymbol{\phi}$ indexes the joint distribution $g(\xi_v, \theta_v)$, which is typically assumed to be multivariate normal with the expected values $E(\xi)$ and $E(\theta)$, the variances Var(ξ) and Var(θ), and the covariance Cov(ξ, θ). Note that the model thus takes the relationship between ξ and θ into account when estimating the parameters of the model. In this way, the person´s tendency to omit an item is considered when drawing inferences on their ability.[1]

So far, the model-based approach has successfully been used for parameter estimation when the missing data process depends on the underlying trait. In a simulation study, Holman and Glas (2005) generated data sets and varied the degree to which a missing value depended on the ability. Adequate item parameter estimates for the incomplete data matrix were obtained when applying their model to account for non-ignorable omissions. Estimating a unidimensional IRT model, in which missing values are simply ignored, yields adequate estimates of the parameters only if the correlation between ability and missing propensity is less than .4. Generally, a higher dependency leads to more bias, and it is found that an increasing number of items can lessen this effect.

The model-based approach allows for testing the ignorability of the missing process when estimating persons´ abilities (e.g., Pohl, Gräfe, & Rose, 2014; Rose et al., 2010). The models allow for the investigation of (a) the extent of non-ignorability, and (b) the consequence of using a unidimensional IRT model in which missing responses are ignored. The extent of non-ignorability is estimated by the size of the relationship between the missing propensity and the ability. If non-ignorability is present in the data, the comparison of parameter estimates between the unidimensional IRT model ignoring missing responses and the model-based approach can be used to evaluate the robustness of the unidimensional IRT model to violations of MAR and distinctness. If differences in parameter estimates are negligible, it is justified to use the simpler and more parsimonious IRT model ignoring

missing responses. This model is much easier to estimate and is also applicable to data with smaller sample sizes. Pohl et al. (2014) and Rose et al. (2010) used competence test data to compare parameter estimates from the model-based approach to account for non-ignorable omissions with those obtained from the unidimensional IRT model ignoring missing responses. They only found minor differences in ability estimates, even though a non-ignorable missing mechanism existed in the data. The violation to ignorability was small, however, and parameters showed robustness to slight violations of ignorability (cf. Holma & Glas, 2005). This would therefore justify the use of the simpler model in scaling the respective competence data.

There might be another explanation for not finding differences in parameter estimates when applying the model-based approach to real data sets. In the simulation study by Holman and Glas (2005), the missing values in the data set were generated according to the same model which later retrieved the unbiased item parameters. However, the missing processes that take place in actual competence test sessions do not necessarily need to occur according to the proposed model. Two assumptions stand out which seem relevant when looking at the occurrence of missing responses in real data sets. One concerns the dimensionality, the other the distribution of the missing propensity. Some indications exist that they might be violated, and thus threaten the applicability of the model-based approach to real data. Neither the plausibility of these assumptions nor the impact of their violations has been investigated so far. Violations of either assumption may result in wrong inferences regarding ability estimates, and might cause the model-based approach to fail in providing unbiased parameter estimates.

As discussed earlier, the propensity to omit items is incorporated as a second dimension, implying that the manifest omission behavior depends on a single underlying latent variable. Lord (1974) describes it as a new trait “representing [the examinee´s] willingness to omit items” (p. 251). One could also plausibly assume that the omission

process is multidimensional. Studies have shown that item format impacts the skipping behavior (Allen, McClellan, & Stoeckel, 2005; Hardt, 2013; Jakwerth et al., 1999; Koretz et al., 1993). Also, item content might influence the omission process in different ways. While some students may be prone to predominantly skip mathematics items containing algebra, others might rather choose to omit geometry items. Thus, the preference of a certain subject matter might lead to different omission mechanisms for different individuals. This queries the unidimensionality assumption of the missing propensity, which so far has not been tested. Ignoring a possible multidimensionality of the missing propensity may lead to biased ability estimates, which, in turn, results in a failure to properly account for the missing data (Rose, 2013).

A second major threat to the adequateness of applying the model-based approach to actual data lies in the distributional assumption of ξ and θ. The use of the marginal item response model for estimating the parameters of interest requires a specification of a density for the latent variables (see, e.g., Adams & Wu, 2007). It is often assumed that the observed data stem from a randomly drawn sample of the population, in which ξ and θ are bivariate normally distributed. However, the distribution of the amount of omissions per person is usually positively skewed (e.g., Duchhardt & Gerdes, 2012; Pohl, Haberkorn, Hardt, & Wiegand, 2012), where many participants omit a few items, and hardly any participants omit many items. As a consequence, the joint distribution of the latent missing propensity and the latent ability may deviate from the bivariate normal. Several simulation studies investigating non-normality of the latent distribution when using MML showed that a violation of the assumed distribution biases parameter estimates (Molenaar, 2007; Stone, 1992; Zwinderman & van der Wollenberg, 1990). Item parameter estimates loose accuracy when the actual underlying distribution is vastly skewed, which especially pertains to items in the more extreme ranges of difficulty (Stone, 1992; Zwinderman & van der Wollenberg, 1990). Furthermore, the recovery of person parameters lacks precision, with, yet again, greater bias

regarding extreme ability levels (Stone, 1992). Both biases decrease with an increasing amount of items, but are still present for item sizes of $I = 20$—a size commonly used in large scale assessments. Considering the response propensity in competence data, the distribution of the amount of omissions is extremely skewed, most β_i are very easy, and most people lie within an extreme level of θ, since they are producing an answer to all or almost all items. Therefore, assuming a normal distribution for θ might pose an actual threat to an application of the model-based approach to actual data.

Due to an incorrect model specification when applying the model-based approach to account for non-ignorable omissions to real data sets, the strengths of the approach as demonstrated in the simulation study by Holman and Glas (2005) might fail to come into display. If the assumptions made do not hold in empirical applications, the model may need to be altered in terms of the assumed dimensionality and the distributional restrictions. If the missing propensity in competence tests is, indeed, multidimensional, a multidimensional model should be used to adequately describe the missing data process. If inaccurate distributional assumptions bias parameters of interest, more general models might be required. The current study aims at verifying or, if necessary, finding alternate specifications for the model-based approach. A model properly accounting for non-ignorable omissions in competence tests making adequate assumptions, is a necessary prerequisite to determine whether the amount of missing values typically observed in large scale studies can be ignored. We specifically test whether unidimensionality of the missing propensity and the distributional assumptions hold, and how existing violations of those assumptions affect ability estimates. The first research question was: Are the assumptions of unidimensionality of the missing propensity and bivariate normal distribution violated, and if so, do these violations have an effect on ability estimates? The second research question dealt with the robustness of the approach ignoring missing responses: Is it necessary to account for non-ignorable missing responses using the model-based approach, or does the simpler model, in

which missing responses are ignored, suffice? Do these results depend on adequate model assumptions of the model-based approach?

So far, large-scale studies did not account for non-ignorable missing responses. This may be justified in light of the previous studies, which found that the inclusion of a missing propensity has no considerable effect on parameter estimates. Using more adequate assumptions, we want to examine these findings more thoroughly. If, with more adequate assumptions, these results can be replicated, the use of the simpler model in which missing responses are ignored would be justified. If, however, parameter estimates change when including a missing propensity, the more complex model-based approach is required.

Method

We used data from the National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011). One main objective of the NEPS is to collect longitudinal data on competence development (Blossfeld, Roßbach, & von Maurice, 2011). For this purpose, tests are developed and repeatedly administered to various age cohorts at various significant educational stages. NEPS focuses on a number of fundamental competence domains, such as handling *information and communication technologies* (ICT; Senkbeil, Ihme, & Wittwer, 2013), *science* (SC; Hahn et al., 2013), *mathematics* (MA; Neumann et al., 2013), and *reading comprehension* (RE; Gehrer, Zimmermann, Artelt, & Weinert, 2013). The assessment of competencies in NEPS mainly relies on the collection of responses participants give to a fixed number of items.

We used competence data from the first and second wave of Starting Cohort 4 (SC4) as well as the second wave of Starting Cohort 6 (SC6). The sample in SC4 consisted of N = 15,239 ninth graders attending regular schools in Germany (Skopek, Pink, & Bela, 2013). The sample in SC6 comprised N = 7,256 adults born between 1944 and 1986 (Skopek, 2013). Both studies were carried out in 2010/2011. For the student sample, the data collection took place in a regular school setting, whereas in the adult sample an interviewer administered the

test booklets in the homes of the participants. The tests were administered in paper and pencil format, and lasted about 30 minutes in each domain. The number of items varied between the domains and the cohorts. In the student sample, 36 items were administered for measuring ICT, 28 for science, 22 for mathematics, and 31 for reading comprehension. In the adult sample, only mathematics and reading comprehension were assessed, with 21 and 30 items, respectively. The response formats included simple multiple choice, complex multiple choice, short-constructed response, and matching tasks. In terms of missing values, a distinction was made between not reached items, invalid answers, omitted items, and indeterminable missing responses. The latter label applies to responses containing more than one kind of missing. On average, students skipped 1.7% of the items in science and reading comprehension, and 3% in ICT and mathematics. In the adult sample, the average number of omissions amounted to 8.9% in mathematics and 5.2% in reading comprehension.

The items of the competence tests were scored either dichotomously or polytomously, depending on the number of subtasks of the item. In accordance to the scaling in NEPS, we used a Partial Credit Model (Masters, 1982) as the basic scaling model, assuming unidimensonality of the latent ability variable (Pohl & Carstensen, 2012). Missing responses in the data were ignored, meaning that they were treated as if the item had not been presented to the examinee. Note that in the models incorporating a missing propensity, the part of the measurement model for the latent ability corresponds to the basic scaling model.

For constructing the missing data indicators, d_{iv} was coded 0 if the answer of person v on item i, $x_{iv,}$ was omitted, 1 if x_{iv} was observed, and 9 otherwise. Due to the fact that a missing value on the last item within a domain is always coded as *not reached*, no omissions were recorded for these items, and the respective missing indicators were excluded from analyses. The missing data indicators of the various competencies therefore consisted of one item less than the number of items in the respective domain.

Only if the missing data are non-ignorable, the model-based approach accounting for non-ignorable omissions by Holman and Glas (2005) is needed. We examined the amount of non-ignorability present in the data by estimating the latent correlation between ability and missing propensity.

Investigating the Appropriateness of the Model Assumptions

Dimensionality.

First, we evaluated whether the assumption of unidimensionality of the missing indicators holds, and whether a violation to that assumption has an effect on ability estimates.

Investigating the dimensionality of the missing propensity.

Testing for unidimensionalty of the missing propensity, we fitted a unidimensional Rasch model to the missing data indicators for each competence domain, using the software ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007). The estimation method was Gauss-Hermite quadrature with 20 nodes for each dimension in reading comprehension. In order to enhance estimation accuracy 25 nodes per dimension were used in ICT and science. We constrained the mean of the latent variable to be zero. The convergence criterion was a .0001 minimum change in deviance. For computational reasons, that is the relatively low amount of missing values on some of the items, which might results in estimation problems, we decided to employ the more restrictive Rasch model as opposed to a two-parameter logistic model (Birnbaum, 1968). In our model, the probability of observing a response is given by

$$p(d_{iv} = 1|\theta_v, \delta_i) = \frac{\exp(\theta_v - \delta_i)}{1+\exp(\theta_v - \delta_i)}. \quad (3)$$

We analyzed *weighted mean squares* (WMNSQ), item characteristic curves, and point-biserial correlations between the number of observed responses and the respective missing data indicator in order to evaluate whether the missing indicators fit the unidimensional Rasch model. To additionally test the assumption that the process underlying the omission behavior

is unidimensional, we compared a *unidimensional missing propensity* (MP1D) model against a *two-dimensional missing propensity* (MP2D) model. In the MP1D model, all missing data indicators load on one latent variable, θ, and the probability of observing an answer from person v on item i is given in Equation 3. As the response format impacts the omission behavior, we allocated the missing data indicators in the MP2D model to two dimensions based on the response format. We distinguished missing data indicators, d_{iv}, of (1) items with multiple-choice format, which constituted the dimension *simple format*, from (2) complex multiple-choice or matching task items,[2] which constituted the dimension *complex format*. Thus, two latent variables, θ_1 and θ_2, were modeled: θ_1 represents a persons´ missing propensity on items with a simple response format; θ_2 represents a persons´ missing propensity on items with a complex response format. In the MP2D model, the model equation of the missing data indicators d_{iv} is

$$p(d_{iv} = 1|\boldsymbol{\theta}_v, \delta_i) = \frac{\exp(\boldsymbol{\theta}_v - \delta_i)}{1+\exp(\boldsymbol{\theta}_v - \delta_i)}, \quad (4)$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The MP2D model was applied to the ICT, science, and reading comprehension data. In mathematics, the number of items which featured more complex formats deemed too small to form a separate dimension, and the dimensionality of the missing propensity was therefore not tested in this domain. In the school sample, the reading domain consisted of 27 items with simple multiple-choice format, and 4 items with more complex formats. In science and ICT, the numbers were 19 and 29 missing indicators for the dimension representing simple response format, respectively, and 9 and 6 for the dimension representing complex response format, respectively. In the reading domain of the adult sample, 23 items constituted the dimension of simple response format, and 7 items the dimension of complex response format. Since the likelihood ratio test is influenced by sample size, we additionally consider Akaike´s information criterion (AIC; Akaike, 1973, 1974), the Bayesian information

criterion (BIC; Schwarz, 1978), and the size of the correlation between the two dimensions when comparing the unidimensional and the two-dimensional models.

Impact of the dimensionality assumption on person parameter estimates.

After having tested for dimensionality of the missing propensity, we investigated the impact of possible violations of the unidimensionality assumption on ability estimates. We estimated ability parameters using the model-based approach to account for non-ignorable omissions (Holman and Glas, 2005). Our ABILITY_MP1D model equals the model-based approach as proposed by Holman and Glas (see Equation 2), where the ability variable is denoted by ξ, and the unidimensional missing propensity is denoted by θ (see Figure 1). Our ABILITY_MP2D model is an extension of this model, in which the missing propensity is modelled as two dimensions. We compared the expected a posteriori (EAP; Mislevy & Stocking, 1989) ability estimates from both models to evaluate the impact of the dimensionality assumption of the missing propensity on ability estimates. This comparison is of particular interest with regard to the domains in which unidimensionality of the missing propensity did not hold. It shows how robust the ability estimates are to violations of the unidimensionality assumption.

Distributional Assumptions.

Second, we evaluated whether the assumption of multivariate normality holds, and whether a violation to that assumption has an effect on ability estimates.

Investigating the distributional assumption.

To answer the second research question regarding the violations of the normal distribution assumption of the missing propensity, several general diagnostic models (GDM; von Davier, 2005a) were fitted using the software *mdltm* (von Davier, 2005b). In the GDM approach, *discrete* latent variables are modeled. An advantage of this includes that the skill distribution can take on various forms and is not restricted to the multivariate normal. Furthermore, the software permits multiple dimensions as well as a combination of

dichotomous and polytomous items. When using discrete latent variables, the GDM takes the form of a located latent class (LLC) model (McCutcheon, 1987; Xu & von Davier, 2008), which departs from the IRT concept which presumes continuous latent variables. Instead, the skill distribution is conceptualized as an ordered set of a finite number of classes *h* (Xu & von Davier, 2008). If a test contains several skill dimensions, the latent classes capture all the realized attribute combinations of the skills, so that the entire discrete latent skill space *P*(*h*) can be represented. Besides the option of estimating a parameter for each of the skill combinations, Xu and von Davier (2008) extended their compensatory GDM by structuring the latent class distribution. They make use of log-linear smoothing (Holland & Thayer, 1987), an approach in which an unsaturated log-linear model preserves fewer characteristics of the observed distribution. The marginal log-likelihood of the structured GDM is given by

$$l = \log L = \sum_{h=1}^{H} n(h) \log P(h) + \sum_{h=1}^{H} \sum_{i=1}^{I} \sum_{k=1}^{K_i} n(i,h,k) \log P(x_i = k \mid h), \quad (5)$$

where *n(h)* captures the number of persons who are in latent class *h*, *i* indexes the items, and K_i denotes the number of response categories for item *i*. $P(x_i = k \mid h)$ is the probability of a person scoring in category *k* on item *i*, given the latent class *h*, and can be modeled using the compensatory GDM (see, e.g., Xu & von Davier, 2008).

In our study, we estimated the ABILITY_MP1D model while making various assumptions for the discrete latent skill space: a saturated model, a structured GDM with a maximum number of 6 moments, and a structured GDM with a maximum number of 2 moments to describe the distribution. The loglinear model describing the latent skill space *P*(*h*), or *P*(ξ, θ), using 2 moments can be written as

$$\log P(\xi, \theta) = \beta_{(0)} + \beta_{(1)}\xi^1 + \beta_{(2)}\xi^2 + \beta_{(3)}\theta^1 + \beta_{(4)}\theta^2 + \beta_{(5)}\xi\theta. \quad (6)$$

Note that when modeling the discrete distribution, the means and variances of the two latent variables as well as their covariance are estimated. This model therefore represents the analog to assuming a bivariate normal distribution (Holland & Thayer, 2000). For the ability dimension, we used 15 skill levels in order to sufficiently reflect the skill space, constraining the attribute space from -2 to 5; for the missing propensity, the attribute space ranged from 2 to 6 with 6 skill levels. Therefore, the maximum number of moments for the missing propensity actually equals five. Due to the limited variance of the missing propensity variable, the fewer number of skill levels for the missing propensity sufficed in order to adequately reflect the skill space. When modeling the latent skill space using 6 moments, 7 additional parameters—four higher order moments for ability and three for missing propensity—were estimated.[3] In sum, the ABILITY_MP1D model was fitted using the three distributional alternatives. The convergence criterion was a .0001 minimum change in deviance. In order to investigate the appropriateness of the distributional restrictions, the models were compared in terms of their deviance, their AIC, and their BIC.

Impact of distributional assumptions on person parameter estimates.

Since one of the main interests of the study was investigating the influence of the distributional assumption on person parameter estimates, we examined how alternate assumptions regarding the joint distribution affect the ability estimates. We therefore compared the EAP ability estimates from the saturated model with EAPs from the models using six and two moments, respectively.

Comparing Person Parameter Estimates from Models with MAR and not MAR Assumptions

In a third step we investigated whether the missing propensity is actually needed in the scaling of competence tests in large-scale assessments, or whether a model ignoring omissions—the model often used in large-scale assessments—is robust to violations of ignorability, and, thus, suffices. The results we obtained from the previous analyses informed

about whether a two-dimensional missing propensity and/or less restrictive distribution assumptions are necessary for applying the model-based approach to actual competence data. We thus specified the model by Holman and Glas (2005) accordingly, contrasting the EAP ability estimates from this adapted model against those from the model ignoring missing responses (IGNORE). The comparison offered information on the adequate treatment of missing data in competence tests. Large discrepancies in parameter estimates would indicate that an inclusion of the missing propensity in the measurement model is necessary.

Results

Across the tested domains, the correlations between ability and missing propensity ranged from $r = .086$ to $r = .524$. More skilled participants tended to omit fewer items, and therefore θ was not independent from ξ. With regard to all data sets, both the MAR and the distinctness assumption were violated. The size of the correlations indicate small to medium violations of ignorability.

Regarding the appropriateness of model assumptions and the necessity to include the latent missing propensity in the model, we found similar results for the various competence domains in both samples. In the following, we illustrate the results on the reading comprehension data of the school sample. The results from the other samples and domains are summarized and discussed briefly in terms of differences and similarities.

Investigating the Appropriateness of the Model Assumptions

Dimensionality.

Investigating the dimensionality of the missing propensity.

Regarding the dimensionality of the missing propensity, we evaluated the item fit of the missing data indicators to a unidimensional Rasch model. For all competence domains in both age groups, the models showed a good fit in terms of the WMNSQ, item characteristic curves, and point-biserial correlations. Results revealed that almost all the misfitting missing data indicators derived from items which contained a response format other than simple

multiple-choice, thus identifying the item format as a possible differentiating factor in terms of the omission behavior. To further investigate whether the unidimensionality assumption holds, we contrasted a one- and a two-dimensional missing propensity model. For the two-dimensional model, we used the response format of the item as the criteria for assigning the missing data indicators to the two dimensions *simple format* and *complex format*. The model comparison between the one- and the two-dimensional model for the missing data indicators in the reading domain of the school sample showed a better model fit for the two-dimensional model compared to the one-dimensional model (see Table 1). The change in deviance was significant and the AIC and BIC values were lower for the two-dimensional model. The latent correlation of $r = .79$ supports the conclusion that the missing propensities for the two kinds of response formats differ and cannot be regarded as a unidimensional latent variable. The other domains show similar results, with an exception for science. Here, the information criteria show inconclusive results, since the AIC favors the two-dimensional model, whereas the BIC favors the unidimensional model (see Table 1). When additionally considering the very high correlation of $r = .96$, the results indicate a unidimensional missing propensity in the science domain.

Impact of the dimensionality assumption on person parameter estimates.

We subsequently tested the robustness of the ability parameters against violations of the unidimensionality assumption. We therefore compared reading comprehension ability estimates of the model-based approach including a unidimensional missing propensity (ABILITY_MP1D) with ability estimates from the model-based approach including a two-dimensional missing propensity (ABILITY_MP2D). Figure 2 shows that including the missing propensity either one- or two-dimensionally makes a small difference for some ability estimates, but the estimates were highly correlated ($r = .998$). Therefore, violations of the unidimensionality assumption had a minor impact on person parameter estimates, and the model assuming a unidimensional missing propensity sufficed.

Table 1

Unidimensional Missing Propensity (MP1D) and Two-Dimensional Missing Propensity (MP2D) Model Fit Statistics

Domain (sample)	Model	AIC	BIC	Deviance	LRT	df	p-value	Corr(θ_1,θ_2)
ICT (school)	MP1D	113128	113403	113056				
	MP2D	111497	111787	111421	1635	2	< .001	.58
Science (school)	MP1D	53517	53731	53461				
	MP2D	53509	53738	53449	12	2	< .005	.96
Reading (school)	MP1D	52490	52727	52428				
	MP2D	51918	52170	51852	574	2	< .001	.79
Reading (adult)	MP1D	34927	35134	34867				
	MP2D	34184	34404	34120	747	2	< .001	.77

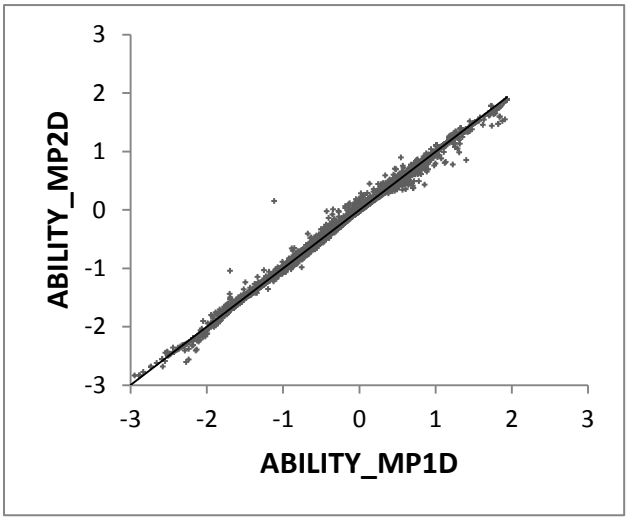

Figure 2. Impact of dimensionality of the missing propensity on ability estimates: Comparing EAP ability estimates from the model-based approach including a unidimensional missing propensity (ABILITY_MP1D) and the model-based approach including a two-dimensional missing propensity (ABILITY_MP2D).

Subsequent analyses showed that highly deviating EAPs stemmed from examinees whose missing propensity on items with multiple-choice format, θ_1, was very different from their missing propensity on items with a more complex format, θ_2. For these individuals, modeling the missing propensity either one- or two-dimensionally made a difference in the estimation of their ability level. As already mentioned, however, this was the case for only a few persons.

Distributional Assumptions.

Investigating the distributional assumption.

To determine the optimal number of parameters needed to describe the joint distribution of ξ and θ, the model fit of the unstructured and the structured GDMs were compared. For the reading comprehension data in the school sample, the AIC favored the saturated model, whereas the BIC preferred the model with six moments (see Table 2). In all other domains except reading comprehension in the adult sample, the BIC as well as the AIC favored the model using 6 moments. In the reading data of the adult sample, the BIC was smallest for the model using only 2 moments.

Table 2

Model Fit Statistics of the Model-Based Approach Including a Unidimensional Missing Propensity (ABILITY_MP1D) With Different Distributional Assumptions

	Domain (sample)											
	ICT (school)		Science (school)		Mathematics (school)		Reading (school)		Mathematics (adult)		Reading (adult)	
Model	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Saturated	781515	782942	619006	620273	451708	452730	423984	425167	144374	145152	167493	168561
6 moments	781460	782300	618932	619611	451573	452122	424607	424607	144356	144700	167403	167941
2 moments	781644	782430	619281	619907	452869	453251	424150	424691	144466	144769	167413	167902

In sum, the multivariate normal distribution did not hold and the 6 moment model best described the data while requiring a more parsimonious number of parameters as compared to the saturated model.

Impact of distributional assumptions on person parameter estimates.

In order to test for the impact of the distributional assumption on person parameter estimates, the EAPs of reading comprehension from the models with two and six moments were compared with the parameters obtained from the saturated model. Figure 3 shows that the EAP estimates differ considerably between models making different distributional assumptions. The results indicate that the use of a model which preserves fewer characteristics of the actually observed joint distribution leads to strongly deviating ability estimates.[4]

a) b)

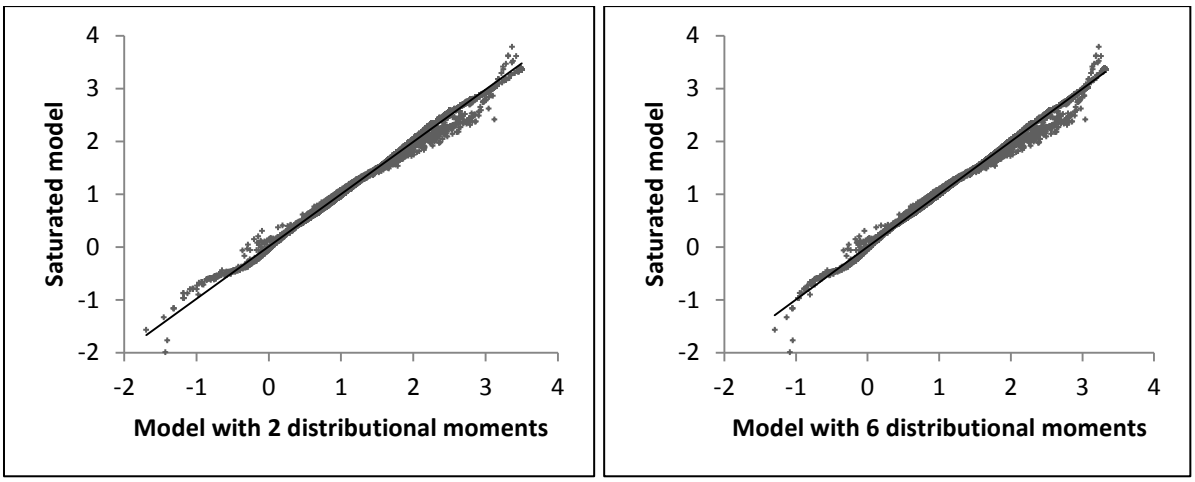

Figure 3. Comparison of ability estimates from the model-based approach including a unidimensional missing propensity (ABILITY_MP1D) for (a) the saturated model and the model using 2 moments, and (b) the saturated model and the model using 6 moments.

For comparisons in all domains and age groups, the correlations between the EAPs from the saturated and the 2 moment models were always smaller than those from the saturated and 6

moment models, indicating that the model using 6 moments better approximated the EAPs from the saturated model than the model using 2 moments. Subsequent analyses showed that EAPs of persons with higher numbers of omitted items were most prone to be affected by distributional assumptions.

Comparing Person Parameter Estimates from Models with MAR and not MAR Assumptions

The previous analyses showed that the ability estimates seemed robust against violations of the unidimensionality of the missing propensity, but not against violations of the bivariate normal distribution assumption. To adequately account for non-ignorable omissions, we therefore decided to use the model-based approach including a unidimensional missing propensity, making no distributional assumptions (i.e., a saturated model). With this model we investigated whether the missing propensity needs to be accounted for or whether the much simpler model in which missing responses are ignored suffices to account for non-ignorable omissions. The comparison of EAP ability estimates for reading comprehension obtained from the model ignoring missing responses (IGNORE) with those obtained from the model-based approach including the missing propensity in the model (ABILITY_MP1D) demonstrates that several ability estimates differed between the two modeling strategies (see Figure 4). Examinees with high ability obtained lower ability estimates when the missing propensity was included in the measurement model, whereas examinees with lower reading competence received higher scores as compared to the model where missing responses were simply ignored. Few persons at the lower (upper) end of the distribution received considerably lower (higher) ability estimates when including a missing propensity. Including the missing propensity in the measurement model for the competence palpably changes the estimated person parameters. The sizes of deviations depend on the ability level. This could be shown for all competence domains and age groups considered in this study.

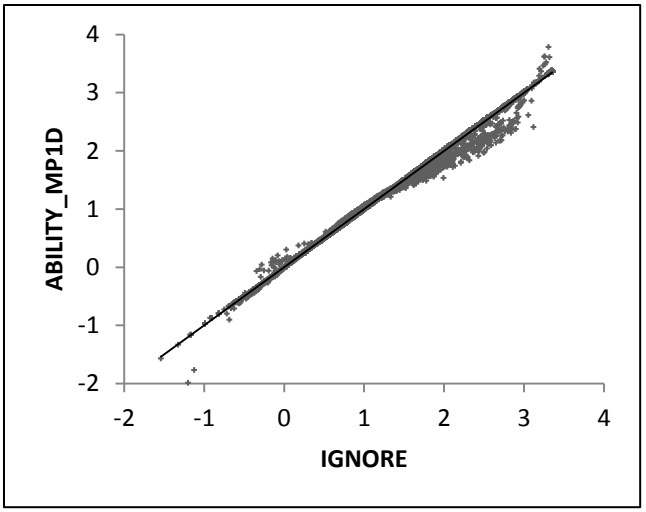

Figure 4. Impact of including the missing propensity in the model: Comparing EAP ability estimates from the unidimensional ability model ignoring missing responses (IGNORE) and the model-based approach including a unidimensional missing propensity (ABILITY_MP1D). In both models, no restrictions are posed on the distribution of the latent variables.

In contrast to results from previous studies which used stricter distributional assumptions, the missing propensity seems to be needed in order to appropriately account for missing responses due to omissions. Information of the missing data indicators is obviously relevant in the scaling; otherwise the ability parameters would not deviate to this extent. When ignoring the prevalent missing data mechanism, non-ignorable omissions are not accounted for, thus resulting in different person parameter estimates.

Discussion

The present study focused on adequately accounting for non-ignorable omissions in competence tests in large-scale assessments. We compared the model by Holman and Glas (2005), which explicitly accounts for a latent missing propensity, to a simpler model in which omissions are ignored. We first investigated the appropriateness of the assumptions made in Holman and Glas´ model. More specifically, we tested the unidimensionality of the missing propensity, as well as the bivariate normal distribution of the missing propensity and the ability. Based on our results, we specified the model using less restrictive assumptions for the

joint distribution, and subsequently tested whether the inclusion of a missing propensity has an effect on ability estimates. The results indicate that although the unidimensionality assumption of the missing propensity did not hold for all considered competence domains, a violation to this assumption had hardly any effect on ability parameter estimates. This justified modeling a unidimensional missing propensity. With regard to the distribution of the latent skill space, the bivariate normal distribution assumption was violated, and the saturated model was used for estimating person ability and missing propensity. The estimated ability parameters from this model deviated from the parameters estimated with a model in which missing values were simply ignored. It can be concluded that a latent missing propensity with an adequate distribution assumption needs to be included in the measurement model of abilities in order to appropriately account for missing responses due to omission.

While previous studies which assumed a bivariate normal distribution found no effect for ability estimates when including the missing propensity (e.g., Pohl et al., 2014; Rose et al., 2010), our results show that when specifying a more flexible distribution, accounting for the latent missing propensity does have an impact on ability estimates, particularly at the upper and lower ends. These findings also concur with previous investigations on the impact of the distribution assumption. Vastly skewed distributions especially introduce bias to person parameter estimates at the ends of the latent continuum (Stone, 1992). These regions of the distribution were precisely the regions where most differences occurred when comparing the model ignoring the missing values and the model-based approach including the missing propensity. When neglecting the skewness of the missing propensity, as was done in previous studies, estimates from the two different scaling models were more alike. The bivariate normal distribution assumption biased the estimates at the ends of the continuum, thus concealing actual existing differences.

In this study we focused on ability estimates on the individual level. In large-scale assessments, however, researchers are usually not interested in individual scores but rather

group statistics (e.g., the relationship between reading ability and gender). The inclusion of the missing propensity in the model will probably have a weaker impact on those group statistics. Opposed to individual person parameters, aggregated group statistics such as means and correlations might prove to be relatively robust to the inclusion of the missing propensity—provided that the group variable is not strongly correlated to the amount of omissions. We conducted exemplary group-level analyses with our data. We first used different models for scaling reading competence, and subsequently performed regression analyses of reading competence on gender. We found no major discrepancies between the estimated regression coefficients or the respective standard errors. This indicates that in practical application, the simpler models ignoring the missing values might suffice. However, this needs further investigation, since we only conducted a single, rather basic analysis. Results might be different for more complex models or models with sub-groups of smaller sample sizes. Our study did show discrepancies in parameter estimates on the individual level, depending on the underlying scaling model. The choice of the scaling model might therefore prove relevant for high-stakes assessment studies, which give feedback to the individual test taker. The individual test score often impacts important decisions such as selection into a certain educational institution. The underlying scaling model should be carefully considered, since it might significantly affect this outcome. Note, however, that our results might not generalize to high-stakes assessments, since the missing data mechanism in these studies deviates from the mechanism in low-stakes assessments. Future research may benefit from applying our methods to examine the applicability of the model-based approach to high-stakes assessment data.

To address some limitations, the current study only focused on intentional omissions while ignoring the missing values that occurred due to time constraints. Some evidence exists in the literature that not-reached items also depend on ability (e.g., Culbertson, 2011) and should be taken into account when estimating competence scores (e.g., Glas & Pimentel,

2008). Recently, further alternatives for dealing with omitted and not-reached items were introduced (Rose, 2013; Rose & von Davier, 2013; Rose, von Davier, & Nagengast, 2013). Rose (2013) proposed joint MIRT models, which consider both types of missing responses simultaneously. Since our major aim lay in investigating the appropriateness of modeling the propensity for omitting an item as proposed in those models, we ignored the not-reached items in our analyses. For the scaling of competence data, however, all missing values should be taken into account.

In the present study, the focus lay on the relationship between the probability for a missing value and the ability of a student. Thus, the non-ignorable missing responses due to a dependency between the missing propensity and ability are taken into account. However, the probability for a missing value in fact depends on other covariates (Köhler, Pohl, & Carstensen, submitted). The missing values could therefore still be non-ignorable with regard to other unobserved variables. In the literature, some models exist which include additional covariates in order to better explain the missing data mechanism (e.g., Moustaki & Knott, 2000; Rose, 2013). The performance of such an approach does depend on the choice of the correct covariates. So far, no study systematically investigated which covariates are relevant for accounting for the missing data mechanism on competence items in large-scale assessments.

When discussing the dimensionality of the missing propensity, only a two-dimensional model based on the response format was specified as an alternative model. However, multidimensionality might still exist. For example, the content area of the item may lead to a different skipping behavior for different people. In further studies, multidimensionality of the missing propensity could be investigated for other aspects. Further note that true values were unknown in our study, and only a comparison between two models with different dimensionality assumptions based on the response format was undertaken. Since the actual values of the complete data matrix remain missing, we have no means to ascertain the

correctness of either of the models. Simulation studies might serve as a basis for investigating the impact of dimensionality assumptions on actual estimation bias (Rose, 2013).

Regarding generalizability, we only used data from one study. As the results were consistent across four domains and two age cohorts, our results most likely generalize to other low-stakes assessments. In contexts with a different missing process, such as high-stakes assessments, different processes than those found in our data may occur. Our procedure of investigating whether or not, and in what form a missing propensity needs to be included in the model may be appropriate for further investigating the ignorablity of missing values in these studies as well. In fact, it would be very interesting to ascertain how the missing behavior and ignorability of omissions differs between low- and high-stakes assessments. Models including a missing propensity may prove valuable for this endeavor.

References

Adams, R. & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.

Adams, R. & Wu, M. (2007). The mixed-coefficients multinomial logit model: A generalized for of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models - Extensions and applications* (pp. 57-75). New York: Springer.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

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

Allen, N. L., McClellan, C. A., & Stoeckel, J. J. (2005). *NAEP 1999 Long-term trend technical analysis report: Three decades of student performance* (NCES 2005–484). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (eds.) (2011). Education as a lifelong process – the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Special Issue 14*.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35*, 321–364. doi: 10.1207/S15327906MBR3503_03.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Culbertson, M. (2011). *Is it wrong? Handling missing responses in IRT.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, USA.

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,* 213-234. doi: 10.1111/j.1745-3984.2001.tb01124.x

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, 39,* 1–39.

Duchhardt, C. & Gerdes, A. (2012): NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225-245. doi:10.1111/j.1745-3984.2008.00062.x

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5*, 50-79.

Glas, C. A. W. & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68,* 907-922. doi: 10.1177/0013164408315262

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I. M., & Prenzel, M. (2013). Assessing scientific literacy over the lifespan - A description of

the NEPS science framework and the test development. *Journal for Educational Research Online, 5*, 110-138.

Hardt, K. (2013). *Using mixed hybrid models to identify testable students with special educational needs in large-scale assessment studies* (unpublished master´s thesis). Otto-Friedrich-University Bamberg, Germany.

Hohensinn, C. & Kubinger, K.D. (2011). Applying Item Response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732- 746. doi: 10.1177/0013164410390032

Holland, P. W. & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Technical Report 87-79). Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183. doi: http://dx.doi.org/10.3102/10769986025002133

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*, 1-17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions*, NAEP Validity Studies, Working Paper Series, American Institutes for Research, Palo Alto, CA.

Johnson, E. G. & Allen, N. L. (1992). *The NAEP 1990 technical report* (Rep. No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Köhler, C., Pohl, S., & Carstensen, C. H. (submitted). *Investigating mechanisms for missing responses in competence tests.*

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Tech. Rep. No. 357). Los Angeles: Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles.

Little, R. & Rubin, D.B. (2002). *Statistical analysis with missing data.* New York: Wiley.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264. doi: 10.1007/BF02291471

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 Technical Report.* Chestnut Hill, MA: Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi: 10.1007/BF02296272

McCutcheon, A. L. (1987). *Latent class analysis.* Thousand Oaks, CA: Sage.

Mislevy, R. J. & Stocking, M. L. (1989) A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75. doi: 10.1177/014662168901300106

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Molenaar, D. (2007). Accounting for non-normality in latent regression models using a cumulative normal selection function. *Measurement and Research Department Reports, 3*. Arnhem: Cito.

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

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, & E. Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online (JERO), 5*, 80-109.

O'Muircheartaigh, C. & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A, 162*, 177-194. doi: 10.1111/1467-985X.00129

Pohl, S. & Carstensen, C. H. (2012). *NEPS technical report - Scaling the data of the competence tests (NEPS Working Paper No. 14).* Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in Item Response Theory models. *Educational and Psychological Measurement, 74*, 423–452. doi: 10.1177/0013164413504926

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement.* Ph.D. thesis, Friedrich-Schiller-University Jena, Dept. of Methodology and Evaluation Research.

Rose, N., & von Davier, M. (2013). *Latent regression and multiple-group IRT models for nonignorable item-nonresponses.* Manuscript in preparation.

Rose, N., von Davier, M., & Nagengast, B. (2013). *Handling of omitted and not-reached items in latent trait models.* Manuscript in preparation.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report ETS RR-10-11), Princeton, NJ: Educational Testing Service.

Rubin, D.B. (1976) Inference and missing data. *Biometrika, 63*, 581-592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association, 95*, 144–154. doi: 10.1080/01621459.2000.10473910

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464. doi: 10.1016/j.stamet.2010.01.003

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The test of technological and information literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online, 5*, 131-169.

Sijtsma, K. & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505–528. doi: 10.1207/s15327906mbr3804_4

Skopek, J. (2013). *Data Manual. Starting cohort 6: Adult education and lifelong learning.* Release 3.0.1. NEPS Research Data Paper. University of Bamberg.

Skopek, J., Pink, S., & Bela, D. (2013). *Data manual. Starting cohort 4: Grade 9 (SC4).* NEPS SC3 1.1.0. NEPS Research Data Paper, University of Bamberg.

Stocking, M. L., Eignor, D., & Cook, L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true score equation procedures* (RR-88-41). Princeton, NJ: Educational Testing Service.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of multilog. *Applied Psychological Measurement, 16*, 1-6. doi: 10.1177/014662169201600101

von Davier, M. (2005a). *A general diagnostic model applied to language testing data* (ETS Technical Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (2013). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. Dordrecht: Springer.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0. Generalised item response modeling software.* Victoria: ACER Press.

Xu, X. & von Davier, M. (2008). *Fitting the structured general diagnostic model for NAEP data* (ETS Research Rep. No. RR-08-27). Princeton, NJ: Educational Testing Service.

Zwinderman, A. H. & van denWollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the rasch model. *Applied Psychological Measurement, 14*, 73-81. doi:10.1177/014662169001400107

Footnotes

[1]Alternate models where the probability for a correct response also depends on θ_v or where the probability for responding to an item also depends on $\xi_v$are possible. All three models, however, can be transformed into each other, and the model in Equation 2 is computationally the simplest and the most straightforward to interpret (Holman & Glas, 2005).

[2]Note that matching items only occur in the domain *reading comprehension*. In the Scientific Use Files (SUF) provided by NEPS, items with complex multiple-choice format are not distinguished from matching task items.

[3]Although 6 moments are modeled with regard to the ability dimension and only 5 moments are modeled with regard to the missing propensity, we will continue referring to this model as the model using 6 moments.

[4]In light of these results, the question arose whether the distributional assumption played a role in not detecting major differences in person parameter estimates in the dimensionality analyses. We therefore reran those models with *mdltm,* using the three distributional alternatives for the ABILITY_MP2D model. When comparing these estimates against those from the ABILITY_MP1D models, the discrepancies remained unobtrusive.

Chapter 3

Study 2: Investigating mechanisms for missing responses in competence tests

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57(4),* 499-522.

3 Investigating mechanisms for missing responses in competence tests

Carmen Köhler[1], Steffi Pohl[2], and Claus H. Carstensen[1]

[1]Otto-Friedrich-University Bamberg, Germany

[2]Free University Berlin, Germany

Author Note

Carmen Köhler, Otto-Friedrich-University Bamberg, Germany; Steffi Pohl, Free University Berlin, Germany; Claus H. Carstensen, Otto-Friedrich-University Bamberg, Germany.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Priority Programme 1646: Education as a Lifelong Process (Grant No. PO 1655/1-1).

This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 3–5th Grade (Paths through Lower Secondary School - Education Pathways of Students in 5th Grade and Higher), doi:10.5157/NEPS:SC3:2.0.0; This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 4–9th Grade (School and Vocational Training - Education Pathways of Students in 9th Grade and Higher), doi:10.5157/NEPS:SC4:1.0.0. This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, the

NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

Correspondence concerning this article should be addressed to Carmen Köhler, Otto-Friedrich-University Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany. Phone: +49-951-863-3406. E-Mail: carmen.koehler@uni-bamberg.de

Abstract

Examinees working on competence tests frequently leave questions unanswered. As the missing values usually violate the *missing at random* condition, they pose a threat to drawing correct inferences about person abilities. In order to account appropriately for missing responses in the scaling of competence data, the mechanism resulting in missing responses needs to be modeled adequately. So far, studies have mainly focused on the evaluation of different approaches accounting for missing responses, making assumptions about the underlying missing mechanism. A deeper understanding of how and why missing responses occur can provide valuable information on the appropriateness of these assumptions. In the current study we investigate whether the missing tendency of a person depends on the competence domain assessed, or whether it can be considered a rather person-specific trait. Furthermore, we examine how missing responses relate to ability and other personality variables. We conduct our analyses separately for not-reached and omitted items, using data from the National Educational Panel Study (NEPS). Based on an IRT approach by Holman and Glas (2005), we investigate the missing process in the competence domains information and communication technologies, science, mathematics, and reading, which were assessed in three age cohorts (fifth-graders: N = 5,193, ninth-graders: N = 15,396, adults: N = 7,256). Results demonstrate that persons' missing propensities may, to some extent, be regarded as person-specific. The occurrence of omissions and not-reached items mainly depends on persons' competencies, and is different for people with a migration background and for students attending different school types, even after controlling for competencies. Our findings should be considered in approaches aiming at accounting for missing responses in the scaling competence data.

Keywords: missing data, missing propensity, Item Response Theory, scaling competencies, large-scale assessment

Theoretical Background

Large-scale assessment studies such as the National Assessment of Educational Progress (NAEP) or the Programme for International Student Assessment (PISA) aim at recording students' learning acquisitions in order to inquire and evaluate the educational system. The employed tests typically assess competencies in areas that are considered important for future success of the individual as well as for the country (e.g., OECD, 2009). To measure competencies, examinees are usually asked to respond to questions, referred to as items. Participants' answers to these items are subsequently scaled using Item Response Theory (IRT) models, drawing inferences on the person's ability level. When working on a test, examinees' occasionally fail at responding to every item presented to them. The occurrence and treatment of these missing values has been widely discussed in literature. Large numbers of missing values pose a threat to the validity of inferences, as the inferences drawn from the incomplete data on, for example, persons' abilities, might deviate from those one would have obtained if the data had been complete (Rubin, 1976).

Although test developers aim at maximizing the response rates in order to decrease uncertainty regarding the validity of the results, missing values still occur. Most prominent among them are those due to *not-reached* and *omitted* items. The former refer to items towards the end of the competence test, which the examinee did not reach as a result of time limits. The latter are intentionally skipped items within the test. The amount of missing values in competence tests is quite remarkable. In the PISA 2000 study, for example, where one testing session contained about 65 items, the average number of omitted and not-reached items was 2.5 and 1, respectively (Adams & Wu, 2002). These numbers varied considerably between states, ranging from only 0.5 and 0.1 in the Netherlands up to 5 and 4.5 in Brazil. In 2009, the average number of missing items was 5 for omitted items and 2 for not-reached items (OECD, 2012). Here the

missing rates were highest for some of the OECD partner countries, with, on average, more than 12 omitted and 2 not reached items in Albania. When looking at the amount of missing values per item in, for example, the 1990 National Assessment of Educational Progress (NAEP) mathematics test, not-reached rates were higher than omission rates. 13 out of the administered 144 items in grade 12 had omission rates above 10%, and not reached rates above 15% (Koretz, Lewis, Skewes-Cox, & Burstein, 1993).

This relatively large amount of missing responses needs to be dealt with in the scaling of competence test data. So far, researchers have not come to a unanimous conclusion on how to best treat missing responses, and miscellaneous studies handle missing values differently. In PISA, as well as in the Third International Mathematics and Science Study (TIMSS; Martin, Gregory, & Stemler, 2000), omitted and not-reached items are ignored when calibrating item parameters, and treated as incorrect when estimating persons' ability scores (Adams & Wu, 2002). Ignoring items means that they are simply dropped from the likelihood when estimating model parameters, treating them as if they had not been administered to the participant. In NAEP missing responses are dealt with equally for both item and person parameter estimation: Not-reached items are ignored, and omitted items are scored as partially correct, with a score corresponding to the reciprocal of the number of options given on a multiple-choice item (Johnson & Allen, 1992).

Each of the aforementioned approaches involves certain assumptions regarding the occurrence of missing responses. Some concerns exist whether these assumptions hold. Treating missing values as incorrect implies that the missing mechanism is purely determined by ability. Furthermore, it presupposes that the participant attempted the item, but could not produce the correct answer. Studies showed, however, that people fail to respond to items for other reasons than lack of knowledge, such as insecurity about the phrasing of the question or lack of

motivation (Jakwerth, Stancavage, & Reed, 1999). This is an argument against treating all missing values as if the participant could not have answered them correctly. The approach of scoring missing values as fractionally correct solves the problem of assuming that an examinee performs worse than guessing, but remains deterministic with regard to the value for the missing response (Rose, 2013). The approach of ignoring missing responses implicitly assumes that the missing mechanism is ignorable (Mislevy & Wu, 1988). According to Rubin (1976), the ignorability assumption holds when the missing data are *missing at random* (MAR), and the parameter vector of the probability density function of the missing-data matrix is distinct from the parameter vector of the probability density function of the complete data matrix. These conditions are usually violated in large-scale assessments. The missing mechanism often depends on the unobserved latent ability, and the parameter vectors are thus not distinct from each other (e.g., Glas & Pimentel, 2008; Holman & Glas, 2005). Overall, violations of the assumptions may lead to biased estimates when applying any of the mentioned approaches.

In an attempt to take the non-ignorable missing mechanism into account, researchers have developed models that include the missing mechanism in the measurement model for ability. The idea behind these model-based approaches is that the missing data holds information on the true distribution of the unobserved latent trait, and should thus be incorporated into the model. Most prominent among the approaches are selection models (Heckman, 1976) and pattern mixture models (Glynn, Laird, & Rubin, 1986; Rubin, 1987). They both attempt at modeling the joint distribution of the missing mechanism and the mechanism for the observed responses, and only differ in their specification of this joint distribution. Selection models and pattern mixture models, however, have their limitations in terms of parameter specification and identification, and are rarely applied in practice. O'Muircheartaigh and Moustaki (1999) have developed the approach of modeling the joint distribution further, using multidimensional IRT models.

Adaptations of their approach resulted in models for omitted (Holman & Glas, 2005) and models for not-reached items (Glas & Pimentel, 2008), as well as in models that simultaneously account for not-reached and omitted items (Rose, 2013). The great contribution and advantage of these model-based approaches over the previously described approaches is that they consider non-ignorability of the missing data. One challenge for these models lies in finding ways of incorporating the missing mechanism in the measurement model. Analog to efforts regarding an adequate scaling model for persons' abilities, the missing mechanism deserves equal consideration in terms of a proper representation. A first step towards establishing how the missing process can be modeled involves determining under which circumstances and for what reasons missing values occur.

In literature some studies exist which investigated reasons for missing responses. Mostly, characteristics of the item such as the difficulty or the response format were examined, showing rather homogeneous findings. Regarding not-reached rates and the influence of the response format, Koretz et al. (1993) found that the first item examinees stop responding to is more likely an item with an open-ended format than a multiple-choice item. In terms of omissions, a similar effect occurs: In the 1990 NAEP study, open-ended questions were the most difficult ones, and also the most likely ones to be skipped (Koretz et al., 1993). A study by Köhler, Pohl, and Carstensen (submitted) additionally showed that the omission behavior differs for multiple-choice items and items with a more complex response format, meaning that the processes leading to an omission on items with different response formats were distinct. Besides the response format, one of the most prominent influencing factors on the omission behavior is the difficulty of the item. Several studies determined that, in general, more difficult items are more frequently skipped (e.g., Koretz et al., 1993; Pohl, Haberkorn, Hardt, & Wiegand, 2012; Rose, von Davier, & Xu, 2010; Zhang, 2013).

Missing values do not solely occur due to specific item or test characteristics, but are also influenced by person characteristics. A number of studies demonstrated that the tendency to respond or not to respond to an item differs between people. These studies mainly deal with omitted items, though some investigated the relationship between the amount of not-reached items and ability. Pohl, Gräfe, and Rose (2014), for example, showed that students with a higher reading ability had higher not-reached rates. These results were, however, not stable across different competence domains. In terms of omitted items, most studies found that more skilled people generally omit fewer items (e.g., Pohl et al., 2014; Rose et al., 2010; Stocking, Eignor, & Cook, 1988; Zhang, 2013). Despite relationships with response format and ability, some studies illustrated differences between omission rates of males and females (e.g., Grandy, 1987; Zhang, 2013), whereas others reported only minor gender discrepancies (Ben-Shakhar & Sinai, 1991; Koretz et al., 1993; von Schrader & Ansley, 2006). Furthermore, Grandy (1987) and Koretz et al. (1993) showed that ethnicity influences the amount of omissions, even after controlling for the proficiency level. A qualitative study by Jakwerth et al. (1999) demonstrated that motivation plays a role in why students omit items, as do test taking strategies and a lack of understanding of the question. Moreover some intercultural differences seem to exist regarding persons' tendencies to omit items (Choppin, 1974; Emenogu & Childs, 2005). Overall, the results indicate that person characteristics do play a role in explaining the tendency to omit and not reach items.

The model-based approaches seem very promising with regard to appropriately accounting for non-ignorable missing values. These models could thus serve as reference models when evaluating different missing data approaches. In order to include the missing mechanism in the measurement model for ability, however, the underlying missing process needs to be known (e.g., Mislevy & Wu, 1996). So far, no information exists on how much of the missing process is inherent in a person, that is, whether it is person-specific. If a person's missing tendency exists as

a construct attributable to the person, it should manifest itself in various testing situations. It might also relate to other constructs or person characteristics, which thus play a role in explaining why missing values occur. If this missing data mechanism is different for various subgroups, these interindividual differences possibly require consideration in the missing data model. The knowledge on how and why missing values occur is necessary in order to establish models which make proper assumptions regarding the missing data mechanism. So far, models including the missing process in the scaling of competence scores have solely incorporated a unidimensional latent omission tendency. If, however, the omission tendency is a rather person inherent construct which relates to other person characteristics, it may be necessary to model the missing data mechanism accordingly. Only a scaling model appropriately including the missing mechanism can adequately account for non-ignorable missing values. Such a model might also serve as a reference model in order to evaluate approaches dealing with missing values differently. While item and test related influences on the occurrence of missing values are quite evident, research on stability of the missing tendency and related other person characteristics is rather inconclusive.

Research Questions

The present study aims at obtaining a comprehensive understanding of the missing data mechanism, that is, the occurrence of missing responses in competence tests of large-scale assessments. We focus on evaluating whether the occurrence of missing responses can actually be attributed to the person. Furthermore, we investigate a broad spectrum of person characteristics that might explain the occurrence of missing responses. Since some studies showed differences between the occurrence of not-reached and omitted items, we examine them separately. We also consider that studies found differences in omissions based on the response format.

If the tendencies to omit and not reach items exist indeed as person-specific constructs, and certain characteristics explain these constructs, the tendencies and their determining factors should be the same across different tests, regardless of test content. It would thus be possible to explain the occurrence of missing values by rather stable, person-specific characteristics. A comprehensive, domain-general model describing the missing data mechanism could be established and incorporated into scaling models for estimating competencies. Such models can provide valuable information on how to best account for non-ignorable missing responses in the scaling of competence tests, since they allow for a comparison to other existing approaches. They can thus aid in determining whether complex models including the missing propensity are necessary, or whether more parsimonious approaches actually suffice.

Our first research question is: To which extent is the occurrence of a missing value person-specific, and therefore not purely determined by characteristics of the item and the tested domain? In other words, do the *missing propensities* for not-reached and omitted items exist as constructs inherent in a person, and can thus predict the response behavior in other situations or tests? We secondly investigate interindividual differences between peoples' missing propensities.

Method

Data

The current study employed data from the National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011). In NEPS, a multi-cohort sequence design serves as the basis for data acquisition on competencies, competence development, and its determining factors. There are six starting cohorts: early childhood, kindergarten, grade 5, grade 9, college students, and adults. The competencies measured involve fundamental domains, for example *information and communication technologies* (ICT; Senkbeil, Ihme, & Wittwer, 2013), *science* (SC; Hahn et al., 2013), *mathematics* (MA; Neumann et al., 2013), and *reading competence* (RE;

Gehrer, Zimmermann, Artelt, & Weinert, 2013), as well as more general, context-free skills, such as *perceptual speed* or *deductive reasoning*. Besides the assessment of competencies, data on relevant background information associated with competence acquisition and progress are also collected.

In order to obtain a general understanding of the occurrence of missing values in a wide age range, we used competence data from three different age cohorts, namely students in grade five (N = 5,193), students in grade nine (N = 15,293), and adults (N = 7,256). The fifth- and ninth-graders attended regular schools in Germany and were, on average, 10.5 (SD = 0.64) and 14.7 (SD = 0.72) years old, respectively. The participants of the adult sample were, on average, 48.3 years old (SD = 10.9). In the student samples, the competence assessment took place in a classroom setting in paper and pencil format. After the testing, which took about two hours, the students answered questions regarding socio demographics, learning environments, attitudes, and further topics. In the adult sample, the assessment took place at the homes of the participants via computer-assisted personal interviewing. After the interviews, which covered schooling, employment, and socio-demographic information, the participants received the competence tests in the form of paper-based booklets. In all cohorts, the randomly administered test booklets differed in sequence of the presented competence tests.

The NEPS competence domains ICT, science, mathematics, and reading were assessed via item sets covering the respective domain. The items were developed in order to fit the Rasch model (Rasch, 1960) or—in case of polytomously scored items—the partial credit model (PCM; Masters, 1982). The tests in ICT, science, and reading mostly contained items with a simple and some items with a complex multiple-choice response format. A simple multiple-choice item consisted of a single question, and required the examinee to choose the correct answer from several presented response options; a complex multiple-choice item entailed several subtasks

(i.e., several questions), each containing two response options. In reading competence, few items were matching tasks, meaning that the examinee was asked to match several headings with corresponding text passages. In mathematics, most items had a simple multiple-choice response format, very few items were complex multiple-choice questions, and some were short constructed response items where the participant was required to insert, for example, the solution to a mathematical problem. In the following, we refer to simple multiple-choice items as having a simple response format and to complex multiple-choice and matching tasks as having a complex response format. All considered samples were tested in mathematics and reading competence. In the ninth-grade sample, additional data was available in the two domains ICT and science. Table 1 gives an overview of the number of administered as well as the average amount of not-reached and omitted items in each cohort and competence domain. Across all domains and all three cohorts, people omitted, on average, between 1.5% and 8% of the administered items. For not-reached items, the numbers ranged between 1.2% and 10.5%. Not-reached rates were especially high in the reading domain in the fifth-grade and the adult sample. In all cohorts, the amount of omissions was higher in mathematics as compared to the amount of not-reached items, whereas the opposite was the case with regard to the reading domain. Overall, the amount of missing values was not negligible, and might therefore need to be considered in the scaling.

Since one aim of our study was to investigate interindividual differences in the missing propensities, we tried to explain the occurrence of not-reached and omitted items via further competencies collected in NEPS, demographic variables, and personality traits. The competencies included *reading speed*, *perceptual speed*, *deductive reasoning*, *procedural metacognition*, and *declarative metacognition.* The reading speed test consisted of 51 sentences making certain statements, which the participant was asked to rate as either true or false (Zimmermann, Gehrer, Artelt, & Weinert, 2012). The test measuring perceptual speed was time-

limited. It comprised 93 items, which required the examinee to match numbers to certain symbols in a correct order. For measuring deductive reasoning, the examinee was presented with 12 matrices items (see Haberkorn & Pohl, 2013), which were developed by Lang and colleagues (Brunner, Lang, & Lüdtke, 2009; Lang, Kamin, Rohr, Stünkel, & Williger, 2012). In all three of the aforementioned tests, the achieved sum score served as the indicator of a participant's skill level. Procedural metacognition was assessed via the examinee's judgment of their own performance in the competence domains (see Lockl, 2013). The number of correctly answered items was subtracted from the number of items the participant estimated as *answered correctly*, and subsequently divided by the number of actual items in the test. The thus calculated percent difference gave information on an examinee's over- or underestimation of their abilities in the respective domain. For declarative metacognition, participants were presented with texts in which certain scenarios concerning school or leisure activities were described (see Lockl, 2012). Several planning, organizing, and resource management strategies were suggested, and the examinee rated them in terms of their usefulness on a four-point rating scale (from 1 = *not useful at all* to 4 = *very useful*). The 69 single evaluations regarding eight different scenarios were compared with expert ratings, and scored as either correct or incorrect. The subsequently calculated mean test score gave information on a person's overall declarative metacognitive skills. The socio-demographic variables we investigated were gender (f*emale* versus *male*), migration background (*yes* versus *no*), and school type. School type was dummy-coded, so that the three dummy-variables indicated whether a person attended (1) lower secondary school, (2) intermediate secondary school, or (3) comprehensive secondary school; upper secondary school served as the reference group. The considered personality traits involved the five NEO-FFI factors and global self-esteem. The NEPS data provided estimated mean scores for *Neuroticism*, *Extraversion*, *Openness*, *Agreeableness*, and *Conscientiousness* from the BFI-10 short version of the NEO-FFI

(Rammstedt & John, 2007). Global self-esteem was available as a sum score, resulting from ratings of ten items tapping the self-esteem construct (e.g., “I feel useless.”) on a five-point rating scale (with 1 = *does not apply* to 5 = *applies completely*). The demographic variables as well as the personality traits were assessed in a questionnaire subsequent to the competence testing.

Analyses

Analyzing the stability of the missing propensities and their relationships to other variables required measurement models that represent a person’s tendency to omit and not reach items, respectively. Note that when speaking of the missing propensities, we refer to both the tendency to not reach and omit items. The modeling and analyses of both tendencies, however, were conducted separately. According to Rose (2013), we computed the sum score of not-reached items for each individual in each of the tested domains in order to represent their tendency to not reach items. For omitted items, Holman and Glas (2005) proposed to model a latent omission tendency. We therefore recoded the original data matrix **X** containing the responses x_{iv} from person v on item i. In the resulting missing data matrix **D**, the omission data indicators d_{iv} were defined as

$$d_{iv} = \begin{cases} 0 & \text{if } x_{iv} \text{ was omitted} \\ 1 & \text{if } x_{iv} \text{ was observed.} \end{cases} \quad (1)$$

In this way, the propensity for an omission can be modeled as a latent variable using an IRT model. When modeling the omission tendency unidimensionally, the probability for observing a response can be expressed via, for example, the Rasch model (Rasch, 1960) as

$$p(d_{iv} = 1|\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1+\exp(\theta_v - \beta_i)}, \quad (2)$$

where θ_v (v = 1,…,n) represents a person's tendency to answer an item, and β_i, (i = 1,…,k) denotes the difficulty of the omission data indicator. In our analyses, the omission data indicators d_{iv} were coded as *not available* if the participant did not reach the corresponding item. Thus, not-reached items were ignored when estimating a person's omission tendency. Previous studies showed that in ICT and reading, the omission tendency is different for items with a simple multiple-choice response format and items with a more complex response format (Köhler et al., submitted). We therefore modeled the omission tendencies in the respective domains two-dimensionally. Omission data indicators from items with simple multiple-choice format load on the first dimension (D1),θ_{1v}, and indicators from complex multiple-choice or matching task items load on the second dimension (D2), θ_{2v}. The model in Equation 2 was therefore extended to a two-dimensional IRT model: The probability for an observation on an item with a simple or a complex response format can be denoted as

$$p(d_{imv} = 1|\theta_{mv}, \beta_i) = \frac{\exp(\theta_{mv}-\beta_i)}{1+\exp(\theta_{mv}-\beta_i)}, \quad (3)$$

where M equaled the number of dimensions, indexed by m = 1,…, M.

Person-specificity of the missing propensities.

In order to test for person-specificity of the missing propensities, we investigated the stability of not reaching and omitting items across different competence domains. The analyses were conducted in all three age cohorts. Note that for the fifth-graders and the adults, only the relationship between the missing propensities in mathematics and reading could be considered, whereas for the ninth-graders, the relationships between the missing propensities in all four competence domains could be examined. To determine the stability of the tendency to not reach

items across competence domains, we computed manifest correlations between the sum scores of not-reached items across different domains; to test the stability of the omission tendency, we estimated latent correlations between the omission tendencies of different domains. As the omission tendencies were modeled two-dimensionally in ICT and reading, the respective between-item-multidimensional IRT model for evaluating the stability of omissions was six-dimensional in grade nine—two dimensions for modeling the omission propensity in ICT and reading, respectively, one dimension in science and mathematics, respectively. In grade five and the adult sample, the models were three-dimensional—one dimension in mathematics and two in reading. High correlations indicate that the missing propensities depend less on the competence domain, but are rather person-specific.

Relations between person characteristics and the missing propensities.

Our second research question dealt with explaining interindividual differences between peoples' missing propensities. We therefore analyzed whether competencies, socio-demographics, and personality traits influence the tendency to not reach or omit items in all three cohorts, respectively. We conducted these analyses exemplary in the reading domain, and validated our findings based on mathematics data. Five multiple linear regression models were used to determine the relationship between the explaining variables and the tendency to not reach items; five multiple latent regression models were estimated with regard to the omission tendency. All models included the respective other missing propensity (i.e., the one not focused on) as an explaining variable, since previous studies showed dependencies between omitted and not-reached items.[1] The remaining explaining variables were added in blocks in a consecutive order. Since the strongest relations were found between the missing propensities and ability, the first model involved competencies. Model 2 additionally comprised socio-demographics. We thirdly included personality traits, since they might explain interindividual differences in the

missing propensities of the examinees beyond what competencies and demographics already elicit. Model 4 involved interactions between the respective other missing propensity and competencies as well as interactions between the ability in the tested domain and other competencies. Model 5 additionally consisted of interactions between the respective other missing propensity and socio-demographics as well as interactions between the competence in the tested domain and socio-demographics. The interactions were added in order to test whether relationships differ for various subgroups. The competencies included were (a) the ability of the respective domain,[2] (b) reading speed, (c) perceptual speed, (d) deductive reasoning, (e) procedural metacognition, and (f) declarative metacognition. Socio-demographics involved gender, migration background, and school type. The personality traits encompassed global self-esteem and the five NEO-FFI factors. Note that not all variables were available in all three data sets. Since no personality tests were administered to the adults, Model 3 was not estimated in the adult sample.

Due to missing values on some of the explaining variables, we imputed them using the R package MICE (van Buuren & Groothuis-Oudshoorn, 2011). The imputation model encompassed all relevant variables of the regression model, including interaction terms, as well as additional predictors explaining other variables or their missing values. The applied imputation methods were *predictive mean matching* for continuous variables, *logistic regression* for binary variables, and the *ordered logit model* for ordered variables with more than two levels. We used passive imputation in order to preserve the relationships of variables included in interaction terms. We chose 20 iterations, producing a single imputed data set. Based on the imputed data set, we estimated the five multiple regression models with the not-reached variable as the dependent variable, and the five multiple latent regression models explaining the omission tendency. Note that in the reading domain, the omission tendency was modeled two-dimensionally.

For all manifest analyses, we used the software R (R Core Team, 2014). All analyses including latent variables were conducted in ConQuest (Wu, Adams, Wilson, & Haldane, 2007).[3]

Results

Person-specificity of the Missing Propensities

The missing propensities for both types of missing values positively correlated across different domains, meaning that people with a higher propensity to omit items in one domain also tended to have more omitted items in the other domains. The same holds for not-reached responses. Regarding the tendency to not reach items, correlations ranged from $r = .19$ to $r = .46$ (see Table 2). A correlation coefficient above $r = .3$ is considered a medium effect (Cohen, 1988). The tendency to not reach items in one domain can therefore be regarded as a relevant predictor for the tendency to not reach items in other competence domains. In the ninth-grade sample, correlations were higher between not-reached rates in ICT, science, and reading, while not-reached rates in mathematics correlated lower with those in the other domains. This means that the tendency to not reach items in mathematics deviates more from the tendencies in the other three domains. This might be due to the fact that not-reached rates in mathematics were noticeably lower, and most people reached the end of the test. The correlations between the tendency to not reach items in mathematics and reading were similar for ninth-graders ($r = .19$) and adults ($r = .21$), but higher for fifth-graders ($r = .37$). This means that fifth-graders who failed to reach the end in the mathematics test tended to also have items missing at the end of the reading test. This relationship was not as strong for ninth-graders and adults.

The latent correlations between the omission tendencies in different domains are presented in Table 3. Within the same competence domain, the omission dimensions correlated rather high in reading (between $r = .76$ and $r = .81$) in all three cohorts, while they correlate somewhat lower in ICT ($r = .58$). The different omission dimensions are thus more distinct from

each other in ICT than in reading. Between different competencies, they were medium to high, ranging from $r = .24$ to $r = .77$. Consequently, a person's omission tendency remained relatively stable across competence domains. In the ninth-grade sample, omission tendencies in ICT, science and mathematics correlated higher amongst each other than with the omission tendency in reading. Thus, the omission tendency in reading deviated more from omission tendencies in the other domains. When comparing the correlations across the cohorts, correlations between the omission tendency in mathematics and the two omission tendencies in reading were similar in all age cohorts. The amount of person-specificity of the omission propensity seemed to be similar in different age cohorts. As expected, the omission tendencies between reading and ICT correlated higher within the same response format.

Overall, these substantial correlations between the missing propensities demonstrate a relatively stable tendency to not reach and omit items across different testing domains, and can therefore be considered person-specific to a certain extent.

Relations between Person Characteristics and the Missing Propensities

We subsequently investigated which person characteristics explain the missing propensities in the reading domain. As is evident in Table 4, the most prominent predictors of the tendency to not reach items across all three cohorts was the first dimension of the omission tendency (D1: omission tendency on items with simple multiple-choice format) and reading speed: Students with more omissions on simple multiple-choice items reached fewer items; students with higher reading speed reached more items.[4] Note that solely the first dimension of the omission tendency served as a relevant predictor, meaning that the two omission tendencies differently relate to the tendency to not reach items. Persons' actual ability in reading was only a meaningful predictor for the tendency to not reach items in fifth grade, where, surprisingly, students reached fewer items when their ability in reading was higher. From the demographic

variables, migration background and school type were relevant in some of the cohorts: Controlling for competencies, ninth-grade students in lower secondary school or in comprehensive secondary school reached fewer items than students in upper secondary school; adults without a migration background reached more items than adults with a migration background. This indicates that the groups differ for reasons other than their actual competence level. None of the personality variables we added in Model 3 further explained variance of the tendency to not reach items. Consequently, personality variables have no explanatory value with regard to the missing process. In Models 4 and 5, many of the included interactions were meaningful predictors of the dependent variable, especially interactions between omission and other competencies in Model 4, and between omission and demographic variables in Model 5. The relationship between the tendency to not reach items and the omission tendency was therefore not unanimous across all competence levels (with respect to the competencies we investigated) and across all subgroups (with respect to the demographic variables we investigated). Overall, the models explained a substantial amount of variance. R-squared ranged between $R^2 = .25$ and $R^2 = .38$. Table 4 also reveals the highly homogeneous findings across the three age cohorts. Not only were the same predictors relevant, but also the direction of the relationship was identical.

Tables 5 and 6 illustrate the results of the latent regression of the omission tendency on the explaining variables. Note that the omission tendency was modeled two-dimensionally, which allowed us to investigate the relationship between the explaining variables and the omission behavior on simple multiple choice items (D1; see Table 5) and the omission behavior on items with a more complex response format separately (D2; see Table 6). For both dimensions, reading competence and reading speed mainly determined the tendency to omit items: Higher competence levels in reading and higher reading speed concurred with fewer omissions. An

additional important variable regarding the tendency to omit D1 items were not- reached items: A higher tendency to not reach items encompassed a higher tendency to omit multiple-choice items. This was in accordance with the above findings, where the tendency to omit on D1 was relevant for predicting the tendency to not reach items. Regarding the omission of items with a complex response format (D2), deductive reasoning was a significant explaining variable: Students with higher deductive reasoning skills rather responded to D2 items. Furthermore, migration background and school type were relevant predictors in some of the cohorts: People without a migration background as well as higher educated people omitted less D2 items even when controlling for competencies. Except global self-esteem, the included personality variables in Model 3 had no explanatory value for the tendency to omit items. For fifth-graders, global self-esteem enhanced the response behavior to D2 items, meaning that fifth-graders with higher self-esteem attempted more items with a complex response format. Note that the two omission dimensions were explained by quite different variables; hence, the process leading to an omission on a simple multiple-choice item is quite distinct from the process leading to an omission on an item with a more complex response format. In Model 4, the regression models for both omission dimensions (D1 and D2) showed significant interactions between the tendency to not reach items and competencies as well as between the ability in reading and competencies. This indicates that the bivariate relationships in models one to three between the tendency to omit and the tendency to not reach items as well as between the tendency to omit and reading ability was quite different depending on the skill level in other competencies. This should be considered when modeling persons' tendencies to omit items. In Model 5, which included interactions between the tendency to not reach items and demographics as well as between reading ability and demographics, only the interaction between ability and lower secondary school served as an additionally relevant predictor in the ninth-grade sample. In all three cohorts, the models explained the omission

tendency to an extensive amount (D1: .1 < R^2 < .45; D2: .2 < R^2< .43).[5] As for the tendency to not reach items, the directions of the relationships were, in general, identical across the different age cohorts.

In sum, a large amount of variance of both missing tendencies could be explained by the included variables. The fact that similar variables equally affect the missing tendencies from people of different age cohorts indicates that the missing data process can be considered rather constant. Besides the generalizability across different cohorts, a second major finding was the generalizability to other competence domains. Regarding the tendency to not reach items in reading, the main explaining variables were the omission tendency on dimension one, reading speed, and, in some cohorts, school type and migration background. In mathematics, also the omission tendency and reading speed served as the most prominent factors. Regarding the omission tendency in reading, the tendency to not reach items, reading speed, and, in some cohorts, reading ability, school type, and migration were important. In mathematics, also the tendency to not reach items, reading speed, and, in some cohorts, the ability in mathematics, procedural metacognition, school type, and gender were relevant predictors. These homogeneous results underline the stability of the missing data process. The significant interaction terms with other competencies and some demographic variables indicate that those characteristics moderate relationships between the tendency to omit and the tendency to not reach items as well as between the ability and the missing propensities. The missing mechanisms therefore cannot be modeled uniformly across subgroups that differ in the respective competencies and demographic variables. Note that results regarding relations with person characteristics and the two omission tendencies, which we segmented based on the response format, frequently deviated from each other, indicating that the tendency to omit on simple multiple-choice items was quite distinct from the tendency to omit on items with a more complex response format.

Discussion

The aim of the present study was to investigate the mechanisms resulting in missing responses in competence tests. We separated missing responses due to omitted items and due to not-reached items, examining whether the tendencies to omit and not reach items exist as person-specific constructs. We further explored a wide range of other person characteristics that might relate to the missing propensities.

Our results demonstrate that a person's missing propensity in one domain relates to the missing propensity in other domains, which allows the conclusion that the missing propensities are to some extent person-specific. We explained interindividual differences in persons' missing propensities to an extensive amount. They were mainly based on the respective other missing propensity, competencies, and demographic variables. In general, people with higher competencies, without a migration background, and in upper secondary school show lower tendencies to omit and to not reach items. In mathematics, females also had a higher omission tendency than males. Some demographic variables were relevant predictors even after controlling for competencies, meaning that additional factors not included in the present study must exist which explain the persisting differences. Some of the explaining characteristics additionally served as moderators between the two missing propensities and between ability and the missing propensities. This indicates that relationships with the missing processes are different for various subgroups, and might need consideration when modeling the missing data mechanisms. We also found that the tendency to omit items with a simple multiple-choice response format and the tendency to omit items with a more complex response format are quite distinct from each other, and relate to different person characteristics.

Overall, our results replicate and enhance previous findings. Several studies indicated that the amount of missing responses depends on the actual ability of a person (e.g., Pohl et al., 2014;

Rose et al., 2010; Stocking et al., 1988). Our study demonstrates that ability rather plays a role in the omission mechanism than in the mechanism for not reaching items. People with lower ability levels generally tend to omit more items. In our study, this relationship was more pronounced in the mathematics test than in the reading test. We also found other, more domain-general competencies, which related to the missing propensities. Especially reading speed emerged as a dominant factor, explaining both the tendency to omit and the tendency to not reach items. It is interesting to note that even after controlling for the actual ability in the tested domain, slower readers reach fewer items at the end of the test and also skip more items throughout the test. This was the case for all cohorts and not exclusively in the reading, but also in the mathematics domain. Speed obviously plays a relevant role even in low stakes assessments, and needs to be considered in the stage of test development as well as in the scaling. In confirmation with past research (Ben-Shakhar & Sinai, 1991; Grandy, 1987; Koretz et al., 1993; Zhang, 2013), we detected mixed results regarding a gender effect. The tendency to omit items was the same for males and females in reading, but not in mathematics. In mathematics, female fifth-graders and adults omitted more items than males, even after controlling for all other competencies. This might be due to gender discrepancies with regard to self-efficacy in mathematics (e.g., Louis & Mistele, 2011; Vermeer, Boekaerts, & Seegers, 2000). Migration background and school type were also relevant predictors in some of the cohorts insofar that people with a migration background and a lower educational level showed higher missing tendencies, even after controlling for all competencies. This indicates that other factors not investigated in the current study might account for differences between these subgroups. People with a migration background and a lower educational level possibly refrain from attempting items with a complex response format because they perceive them as more difficult. The factors possibly explaining differences between the aforementioned subgroups certainly need further investigation. The

differences should also be considered in the stage of item calibration in order to avoid systematic disadvantages for certain subgroups. In terms of the personality traits examined in our study, none additionally explained participants' missing propensities. They can therefore be disregarded with respect to the missing data mechanism. Various interactions we investigated were relevant, especially those concerning competencies. They serve as moderators between the two missing propensities as well as between ability and the missing propensities, and might need consideration when modeling the missing data mechanism. Lastly, the omission tendencies seemed fairly distinct from each other, and related to different characteristics. These results clearly indicate that missing values on simple multiple-choice items result from a different mechanism than missing values on items with a more complex response format. Since these omission processes differ, they should be handled separately when modeling the missing data mechanism.

In terms of generalizability of the results, we focused on omissions and not-reached items in a low-stakes assessment. In high-stakes assessments, other test-taking strategies might prevail, thus resulting in different missing data mechanisms. Within the framework of low-stakes assessments, however, we could demonstrate person-specificity of the missing propensities in three cohorts with a wide age range. Furthermore, we examined interindividual differences between persons' missing propensities, and showed that, across different age cohorts and two different test domains, the missing propensities equally relate to other characteristics. These results indicate that the missing propensities might be some sort of a construct inherent in a person. According to Cronbach and Meehl (1955), the process of validation involves various inquiries as well as evidence from different sources. Both the stability over occasions and the uniform relationship to other stable person characteristics meet two of the criteria in the validation procedure (Cronbach & Meehl, 1955). Additional indications would be necessary in

order to truly validate the missing propensities as constructs, for example by examining persons' missing propensities across various time points using longitudinal data. This would further verify the stability of persons' missing propensities. Although we were able to identify person characteristics that well predicted the missing propensities, some of the variance between peoples' omission and not-reached tendencies was left unexplained. Future research might consider other possible influences. Motivation, for example, plays a role in the performance on low-stakes tests (Wise & DeMars, 2005), also affecting the amount of omissions (Jakwerth et al., 1999). As the missing propensities were, in part, specific to the tested domain, it would be valuable to investigate further domain related characteristics, such as the self-concept in the respective domain or the fear of failure.

One strength of this investigation is that we integrated research from previous findings, covering a broad spectrum of aspects potentially relevant for explaining the missing propensities. Thus, we identified factors which remain meaningful even after controlling for all others. We were also able to determine some competencies which moderate the relationship between omitted and not-reached items as well as between the missing propensities and ability, and which should therefore be taken into consideration when accounting for missing values. A further novelty of our study was the separation of the omission tendency based on response format. Most large-scale studies make use of several types of response formats, and need to consider that the missing mechanisms differ accordingly. In light of the scaling of competencies and models which aim at including the missing data mechanism in the measurement model, our results demonstrate which variables are relevant in predicting a missing value. The stability of our results further demonstrates that the missing data mechanism is relatively uniform and may be modelled equally across different domains and cohorts. Including the missing propensity as well as relevant variables in the measurement model for ability might enhance the accuracy of parameter

estimates, since such a model can adequately account for non-ignorable missing values. Whether or not such complex models are actually necessary needs to be investigated in future studies. Simulation studies might aid in evaluating to what extent an inclusion of the missing propensity or relevant covariates can improve parameter estimates. However, our results allow assessing some of the assumptions of other approaches. The fact that the probability for a missing value does not solely depend on the ability in the tested domain refutes the assumption that missing values merely result from lack of knowledge. In low-stakes assessments, missing values should therefore not be treated as wrong. Since the missing mechanism differs for various subgroups and is also different with respect to item format, the assumption of a uniform missing mechanism across all persons and all items does not hold, either.

The current study certainly identified relevant aspects of persons' missing tendencies. These should be considered in other studies that aim at modeling the missing data mechanism. Only a model making proper assumptions regarding the missing data mechanism allows drawing adequate conclusions on the influence of non-ignorable missing values on true parameter estimates. Such a model can aid in determining how to account accurately for non-ignorable missing responses.

References

Adams, R. J.,Wilson, M., &Wang,W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. doi: 10.1177/0146621697211001

Adams, R. & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 77–92. DOI: 10.1111/j.1745-3984.1991.tb00341.x

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (eds.) (2011). Education as a lifelong process – the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Special Issue 14*.

Brunner, M., Lang, F. R., & Lüdtke, O. (2009). *Expertise: Erfassung der fluiden Intelligenz über die Lebensspanne im Rahmen der National Educational Panel Study*.

Choppin, B. H. (1974). *The correction for guessing on objective tests* (IEA Monograph Studies, No. 4). Stockholm, Sweden: The International Association for the Evaluation of Educational Achievement.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. 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*, 1-17.

Emenogu, B. C., & Childs, R. A. (2005). Curriculum, translation, and differential functioning of geometry items. *Canadian Journal of Education, 28*, 123–142.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples*. New York, NJ: Springer. doi: 10.1007/978-1-4612-4976-4_10

Grandy, J. 1987. *Characteristics of examinees who leave questions unanswered on the GRE General Test under rights-only scoring. (GRE Board Professional Report No. 83-16P).* Princeton, NJ: Educational Testing Service.

Haberkorn, K. & Pohl, S. (2013). *Cognitive basic skills – Data in the Scientific Use File.* Bamberg: University of Bamberg, National Educational Panel Study.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, limited dependent variables, and simple estimator for such models. *Annals of Economic and Social Measurement, 5*, 475-492.

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*, 1-17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions*, NAEP Validity Studies, Working Paper Series, American Institutes for Research, Palo Alto, CA.

Johnson, E. G. & Allen, N. L. (1992). *The NAEP 1990 technical report* (Rep. No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Köhler, C., Pohl, S., & Carstensen, C. H. (submitted). *Taking the missing propensity into account when estimating competence scores—Evaluation of IRT models for non-ignorable omissions.*

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Tech. Rep. No. 357). Los Angeles: Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles.

Lang, F. R., Kamin S., Rohr M., Stünkel C., & Williger B. (2012). *Abschlussbericht zur Ergänzungsstudie "Erfassung der fluiden Intelligenz über die Lebensspanne im Rahmen der National Educational Panel Study"*.

Lockl, K. (2012): *Assessment of declarative metacognition: Starting Cohort 4 – Ninth Grade.* Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Lockl, K. (2013). *Assessment of procedural metacognition: Scientific Use File 2013.* Bamberg: University of Bamberg, National Educational Panel Study.

Louis, R., & Mistele, J. (2011): The differences in scores and self-efficacy by student gender in mathematics and science. *International Journal of Science and Mathematics Education, 10*, 1163-1190. doi: 10.1007/s10763-011-9325-9

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 Technical Report.* Chestnut Hill, MA: Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

OECD (2009). *Pisa 2006 Technical Report*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 Technical Report, PISA*, OECD Publishing.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in Item Response Theory models. *Educational and Psychological Measurement, 74*, 423–452. doi: 10.1177/0013164413504926

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 15).* Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five inventory in English and German. *Journal of Research in personality,41,* 203–212. doi:10.1016/j.jrp.2006.02.001

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Rose, N. (2013). *Item nonresponses in educational and psychological measurement.* Ph.D. thesis, Friedrich-Schiller-University Jena, Dept. of Methodology and Evaluation Research.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report ETS RR-10-11), Princeton, NJ: Educational Testing Service.

Rubin, D.B. (1976) Inference and missing data. *Biometrika, 63*, 581-592. doi: 10.1093/biomet/63.3.581

Stocking, M. L., Eignor, D., & Cook, L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true score equation procedures* (RR-88-41). Princeton, NJ: Educational Testing Service.

van Buuren, S. & Goothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67.

Vermeer, H. J., Boekaerts, M., & Seegers, G. (2000). Motivational and gender differences: Sixth-grade students' mathematical problem-solving behavior. *Journal of Educational Psychology*, *92*, 308-315. doi: 10.1037/0022-0663.92.2.308

von Schrader, S. & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education, 19*, 41–65.

Warm, T.A. (1989). *Weighted likelihood estimation of ability in item response theory. Psychometrika*, *54*, 427-450. doi: 10.1007/BF02294627

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17. doi: 10.1207/s15326977ea1001_1

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0. Generalised item response modeling software.* Victoria: ACER Press.

Zhang, J. (2013). *Relationships between missing response and skill mastery profiles of cognitive diagnostic assessment.* Ph.D. thesis, University of Toronto, Dep. of Curriculum, Teaching, and Learning.

Zimmermann, S., Gehrer, K., Artelt, C., & Weinert, S. (2012). *The assessment of reading speed in grade 5 and grade 9.* Bamberg: University of Bamberg, National Educational Panel Study.

Footnotes

[1]In order to include the omission propensity as an explaining variable, we used the manifest Weighted Likelihood Estimates (WLE; Warm, 1989) estimated from the latent omission propensity model. In reading, each examinee obtained a score for the omission propensity on simple multiple-choice items, and one for the omission propensity on items with a more complex response format. Both were included in the regression models.

[2]As for the omission propensity, the ability in reading was included using manifest WLE estimates.

[3] The Mixed Coefficients Multinomial Logit Model (MCMLM) fitted by ConQuest is a Rasch-type item response model, including a variety of item response and latent regression models (Adams, Wilson, & Wang, 1997).

[4]Bear in mind that due to the coding of the omission data indicators (see Equation 1) higher values on the omission propensity indicate lower omission rates.

[5]Due to computational errors regarding the estimated latent conditional variance, R-squared should not be compared across the five models.

Table 1

Average amount of missing items per person in IRT-scaled competence tests in three NEPS cohorts

Cohort	Domain	Items	omitted	not-reached
Fifth-graders	*Mathematics* (MA)	24	5.1%	1.2%
	Reading (RE)	32	4.4%	10.5%
Ninth-graders	*Information and Communication*			
	Technologies (ICT)	36	3.5%	4.7%
	Science (SC)	28	1.6%	6.2%
	Mathematics (MA)	22	2.7%	0.6%
	Reading (RE)	31	1.5%	4.6%
Adults	*Mathematics* (MA)	21	8.0%	5.4%
	Reading (RE)	30	3.6%	10.1%

Table 2

Manifest correlations between not-reached tendencies of different domains in three cohorts

	Fifth-graders	Ninth-graders			Adults
Domain	Mathematics	ICT	Science	Mathematics	Mathematics
Science		.46			
Mathematics		.30	.28		
Reading	.37	.36	.39	.19	.21

Table 3

Latent correlations between omission tendencies of different domains in three cohorts

Domain	Fifth-graders		Ninth-graders					Adults	
	Mathe-matics	Reading D1	ICT D1	ICT D2	Science	Mathe-matics	Reading D1	Mathe-matics	Reading D1
ICT D2			.58						
Science			.77	.43					
Mathematics			.70	.47	.70				
Reading D1	.54		.41	.24	.46	.41		.47	
Reading D2	.59	.81	.37	.44	.39	.42	.76	.58	.76

Note. D1 = latent omission tendency on simple multiple-choice items; D2 = latent omission tendency on complex multiple choice or matching task items.

Table 4

Standardized regression coefficients of multiple regressions to predict tendency to not reach items in reading

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Missing propensity															
Omission D1	**-0.29**	**-0.50**	**-0.11**	**-0.29**	**-0.48**	**-0.10**	**-0.29**	**-0.48**		**-0.52**	**-1.13**	-0.10	**-0.44**	**-0.57**	0.02
Omission D2	-0.01	-0.01	**-0.11**	-0.01	-0.01	-0.09	-0.01	-0.01		0.00	0.04	**-0.15**	-0.04	0.06	**-0.26**
Competencies															
Reading competence	**0.17**	0.05	-0.08	**0.20**	0.09	-0.04	**0.20**	0.09		**0.75**	**0.46**	-0.05	**0.42**	**0.26**	-0.06
Reading speed	**-0.32**	**-0.17**	**-0.44**	**-0.31**	**-0.14**	**-0.42**	**-0.31**	**-0.14**		**-0.16**	0.01	**-0.40**	**-0.17**	-0.08	**-0.42**
Procedural metacognition	0.06	0.03	**-0.16**	0.06	0.03	**-0.16**	0.06	0.03		**0.10**	-0.07	-0.03	**0.10**	-0.05	-0.03
Demographics															
Migration background (yes vs. no[b])				0.01	0.03	**0.11**	0.01	0.03		0.01	0.03	**0.11**	-0.01	0.04	**-0.11**
Lower vs. upper secondary school[b]				0.09	**0.18**	0.10	0.09	**0.18**		0.10	**0.17**	0.10	**0.18**	**-0.28**	**0.13**
Comprehensive vs. upper secondary school[b]				0.10	**0.11**	NA	0.09	**0.11**		0.09	0.10	NA	0.03	**-0.26**	NA
Personality															

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Interactions between															
- omission and competencies															
- ability and competencies															
Omission D1 x ability										-0.05	-0.04	**0.16**	-0.06	-0.10	**0.12**
Omission D1 x reading speed										**0.34**	**0.43**	0.01	**0.33**	**0.27**	-0.04
Omission D1 x perceptual speed										**0.11**	**0.28**	NA	**0.11**	**0.26**	NA
Omission D2 x deductive reasoning										0.02	**0.14**	NA	0.01	**0.15**	NA
Omission D2 x procedural metacognition										-0.02	-0.02	0.00	-0.01	-0.02	0.00
Omission D2 x declarative metacognition										0.04	-0.06	NA	0.07	-0.04	NA
Ability x deductive reasoning										**-0.12**	-0.08	NA	-0.06	-0.05	NA
Ability x declarative metacognition										**-0.42**	-0.15	NA	**-0.28**	**-0.13**	NA
Interactions between															
- omission and demographics															
- ability and demographics															

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Omission D1 x gender													0.01	**0.15**	**-0.13**
Omission D1 x migration background													-0.04	-0.08	**-0.31**
Omission D1 x lower secondary school													-0.02	**-0.41**	-0.05
Omission D1 x comprehensive secondary school													-0.04	**-0.24**	NA
Omission D2 x comprehensive secondary school													0.02	-0.11	0.03
Ability x lower secondary school													**0.18**	0.04	0.01
R^2	.246	.332	.348	.258	.350	.365	.258	.352		.293	.379	.370	.304	.340	.380

Note. Standardized regression coefficients with $\beta > .1$ and $p < .05$ are in boldface. SC3 = starting cohort 3 (fifth-graders); SC4 = starting cohort 4 (ninth-graders); SC6 = starting cohort 6 (adults); D1 = latent omission tendency on simple multiple-choice items; D2 = latent omission tendency on complex multiple choice or matching task items, NA = variable not available in data set.

[a]Only variables listed where, in any of the regression models, the standardized regression coefficient was $\beta > .1$ and $p < .05$

[b]Serves as the respective reference group ($x = 0$)

Table 5

Standardized regression coefficients of multiple latent regressions to predict omission tendency on simple multiple-choice items in reading

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Missing propensity															
Not-reached	**-0.15**	**-0.1**	0.01	**-0.14**	-0.09	-0.06	**-0.13**	-0.09		0.07	-0.03	-0.01	0.04	**-0.17**	0.06
Competencies															
Reading competence	-0.02	0.06	**0.15**	-0.02	0.05	**0.12**	-0.03	0.04		**-0.37**	**-0.59**	**0.30**	**-0.40**	**-0.38**	**0.32**
Reading speed	**0.16**	**0.17**	**0.20**	**0.17**	**0.17**	**0.19**	**0.16**	**0.17**		**0.16**	**0.20**	**0.20**	**0.16**	**0.19**	**0.19**
Deductive reasoning	0.03	-0.01	NA	0.02	-0.03	NA	0.02	-0.03		0.03	-0.02	NA	0.03	-0.02	NA
Demographics															
Migration background (yes vs. no[b])				0.01	-0.06	NA	0.01	-0.06		0.01	-0.06	0.02	0.01	-0.06	0.01
Lower vs. upper secondary school[b]				-0.01	-0.04	NA	0.00	-0.05		-0.02	-0.03	-0.08	-0.02	-0.05	**-0.12**
Personality															
Global self-esteem							0.09	-0.01	NA	**0.10**	-0.01	NA	0.10	0.00	NA

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Interactions between															
- not-reached and competencies															
- ability and competencies															
Not-reached x reading speed										0.00	**-0.22**	0.00	0.00	**-0.22**	0.00
Not-reached x deductive reasoning										-0.02	0.00	NA	-0.02	0.01	NA
Not-reached x declarative metacognition										**-0.16**	**0.18**	NA	**-0.16**	**0.20**	NA
Ability x reading speed										-0.03	**0.26**	**-0.16**	-0.02	**0.23**	**-0.18**
Ability x perceptual speed										0.03	-0.03	NA	0.03	-0.03	NA
Ability x deductive reasoning										**0.12**	0.07	NA	**0.12**	0.04	NA
Ability x procedural metacognition										**-0.10**	-0.03	-0.02	**-0.11**	-0.02	-0.02
Ability x declarative metacognition										**0.18**	**0.27**	NA	**0.22**	**0.21**	NA
Interactions between															
- not-reached and demographics															
- ability and demographics															
Ability x lower secondary school													0.01	-0.08	-0.04
R^2	.105	.341	.446	.112	.345	.426	.164	.383		.259	.416	.410	.258	.408	.431

Note. Standardized regression coefficients with $\beta > .1$ and $p < .05$ are in boldface. SC3 = starting cohort 3 (fifth-graders); SC4 = starting cohort 4 (ninth-graders); SC6 = starting cohort 6 (adults); D1 = latent omission tendency on simple multiple-choice items; D2 = latent omission tendency on complex multiple choice or matching task items, NA = variable not available in data set.

[a]Only variables listed where, in any of the regression models, the standardized regression coefficient was $\beta > .1$ and $p < .05$

[b]Serves as the respective reference group ($x = 0$)

Table 6

Standardized regression coefficients of multiple latent regressions to predict omission tendency on items with a complex response format in reading

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Missing propensity															
Not-reached	-0.08	-0.04	-0.06	-0.08	-0.03	-0.05	-0.07	-0.03		-0.15	-0.12	**-0.11**	**-0.20**	-0.08	-0.07
Competencies															
Reading competence	0.06	**0.16**	**0.42**	0.04	**0.14**	**0.36**	0.03	**0.14**		-0.10	0.06	**0.39**	0.02	**-0.25**	**0.37**
Reading speed	**0.12**	0.08	**0.19**	**0.12**	0.07	**0.20**	**0.12**	0.08		**0.11**	0.10	**0.20**	**0.11**	**0.11**	**0.20**
Deductive reasoning	**0.21**	**0.14**	NA	**0.20**	**0.13**	NA	**0.20**	**0.12**		**0.17**	0.08	NA	**0.16**	0.09	NA
Demographics															
Migration background (yes vs. no[b])				-0.04	**-0.12**	0.04	-0.04	**-0.12**		-0.04	**-0.11**	0.04	-0.04	**-0.14**	0.03
Lower vs. upper secondary school[b]				-0.07	-0.06	**-0.13**	-0.06	-0.07		-0.07	-0.08	**-0.13**	**-0.12**	-0.05	**-0.19**
Personality															
Global self-esteem							**0.12**	0.03	NA	**0.13**	**0.03**	NA	**0.13**	0.03	NA
Interactions between															
- not-reached and competencies															
- ability and competencies															

Predictors[a]	Modell 1			Modell 2			Modell 3			Modell 4			Modell 5		
	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6	SC3	SC4	SC6
Not-reached x reading speed										0.09	-0.07	0.00	0.09	-0.07	0.00
Not-reached x deductive reasoning										**0.11**	**0.21**	NA	**0.11**	**0.24**	NA
Not-reached x declarative metacognition										-0.03	0.03	NA	0.00	0.05	NA
Ability x reading speed										0.02	**0.14**	0.03	0.00	**0.20**	0.00
Ability x perceptual speed										-0.05	**0.15**	NA	-0.05	**0.15**	NA
Ability x deductive reasoning										0.04	-0.09	NA	0.00	-0.02	NA
Ability x procedural metacognition										**-0.15**	-0.07	-0.02	**-0.14**	-0.08	-0.02
Ability x declarative metacognition										0.11	**-0.13**	NA	0.09	-0.07	NA
Interactions between															
- *not-reached and demographics*															
- *ability and demographics*															
Ability x lower secondary school													-0.08	**0.13**	-0.03
R^2	.193	.404	.396	.204	.371	.415	.228	.373		.261	.342	.419	.266	.368	.432

Note. Standardized regression coefficients with $\beta > .1$ and $p < .05$ are in boldface. SC3 = starting cohort 3 (fifth-graders); SC4 = starting cohort 4 (ninth-graders); SC6 = starting cohort 6 (adults); D1 = latent omission tendency on simple multiple-choice items; D2 = latent omission tendency on complex multiple choice or matching task items, NA = variable not available in data set.

[a]Only variables listed where, in any of the regression models, the standardized regression coefficient was $\beta > .1$ and $p < .05$

[b]Serves as the respective reference group ($x = 0$)

Chapter 4

Study 3: Performance of missing data approaches in retrieving group-level parameters

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). *Performance of missing data approaches in retrieving group-level parameters.* A modified version of this manuscript has been submitted for publication in the Journal of Educational Measurement.

4 Performance of missing data approaches in retrieving group-level parameters

Carmen Köhler[1], Steffi Pohl[2], and Claus H. Carstensen[1]

[1]Otto-Friedrich-University Bamberg, Germany

[2]Freie Universität Berlin, Germany

Author Note

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Priority Programme 1646: Education as a Lifelong Process (Grant No. PO 1655/1-1).

This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

Correspondence concerning this article should be addressed to Carmen Köhler, Otto-Friedrich-University Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany. Phone: +49-951-863-3406. E-Mail: carmen.koehler@uni-bamberg.de

Abstract

Competence data of low-stakes large-scale assessment studies allow for evaluating relationships on the group-level (e.g., between an explanatory variable and ability). The impact of item-level nonresponse has not been investigated with regard to group-level statistics (e.g., regression coefficients). Methods for dealing with missing values include classical approaches such as treating missing values as incorrect or ignoring them, as well as recent model-based approaches that account for nonignorable nonresponse. We simulate data according to various missing data mechanisms, and study how the missing data approaches perform in retrieving parameter estimates on the group-level. Results show that model-based approaches and ignoring missing values outperformed treating missing values as incorrect responses. The relevance of the results is demonstrated in an empirical example.

Keywords: missing data, missing propensity, Item Response Theory, large-scale assessment, competence test, simulation

Theoretical Background

The goal of conducting large-scale studies is to permit the assessment of learning progresses, to evaluate educational systems, and to address impacts of educational backgrounds on academic careers. Results from large-scale assessments can have a major impact on policies, since political choices in, for example, the educational system might be based on the respective outcomes. The main aim of large-scale assessments therefore lies in an accurate, coherent measurement of competencies. Coherency is necessary in order to allow for objective and fair comparisons between subgroups and across different measurement occasions. The framework of Item Response Theory (IRT) provides a basis for such comparisons, and is typically the method of choice developing and scaling competence tests. In accordance with IRT, competence of an individual in a specific area at a certain point in time is usually assessed via a certain amount of questions, or *items*. One problem that arises in the scaling of the responses stemming from competence tests is that examinees typically fail to respond to some of the items presented to them. When different test takers respond to different items, the item set from which inferences can be drawn varies between persons. This is generally not a problem in IRT, since examinees proficiencies can be estimated irrespective of which items have been presented to them. However, missing values in competence tests might occur according to a certain mechanism. For example, a certain group of examinees might only skip items that were too difficult for them. The process leading to missing values is thus systematic in this group. The process leading to a missing value might be different for another group of examinees, and—depending on how missing values are treated—might impede coherent competence measurement across the two groups.

Many papers have focused on identifying the circumstances under which missing values threaten the accuracy of measurement (Lord, 1974; Mislevy & Wu, 1996; Rubin, 1976).

According to Rubin's (1976) fundamental article, missing values are ignorable when (a) the missing data are either missing completely at random (MCAR) or missing at random (MAR), and when (b) the parameter vector of the probability density function of the missing-data matrix is distinct from the parameter vector of the probability density function of the complete data matrix (Rubin, 1976). The MCAR condition is met if the occurrence of missing values neither depends on any of the observed variables nor on any unobserved variables (Rubin, 1976). The MAR condition is met if the missing data depends on (some of) the observed data—and this relationship is known and can thus be taken into account—but not on any unobserved variables (Rubin, 1976). In IRT context, distinctness refers to independence between the parameter space of the latent trait—for example ability—and the parameter space of the missing data process (Mislevy & Wu, 1996). If the conditions (a) and (b) are met, the missing data can be ignored, meaning that they are treated as if they had not been administered to the examinee. If the conditions (a) and (b) are not met and the missing data process depends on, for example, the unobserved responses, missing values are nonignorable. Ignoring them can lead to incorrect inferences about the latent trait (Mislevy & Wu, 1996).

Whether missing responses are ignorable needs to be assessed separately for different types of missing responses in the respective context they occur. *Omitted items* and *not-reached items* received the most attention in missing data research, since they are typically MNAR, and thus pose the largest challenge on scaling (Mislevy & Wu, 1996). The former refer to skipped items and can appear at each section of the test, whereas the latter typically refer to all missing values after the last valid given response (see, e.g., Lord, 1974). Omitted and not-reached items relate to person characteristics such as gender, ethnicity, and other cognitive abilities (Köhler, Pohl, & Carstensen, in press; Koretz, Lewis, Skewes-Cox, & Burstein, 1993). With regard to these variables, the missing values are MAR, and the variables should be included as covariates

in the measurement model for ability. However, missing values commonly also depend on the ability of the person (Pohl, Gräfe, & Rose, 2014; Rose, von Davier, & Xu, 2010). According to Rubin (1976), they are therefore nonignorable.

So far, large-scale assessment studies employ missing data approaches that do not consider the dependency between missing values and the unobserved responses. Interestingly, the studies differ in how they handle missing data, though they use similar IRT models in the scaling (e.g., Adams & Wu, 2002; Allen, Donoghue, & Schoeps, 2001; Martin, Gregory, & Stemler, 2000; Pohl & Carstensen, 2012). The PISA study and the Third International Mathematics and Science Study (TIMSS) use a two-step procedure. For the international calibration of item difficulties, not-reached items are ignored—meaning they are treated as not administered—whereas omitted items are treated as incorrect; for the student score generation, omitted and not-reached items are both treated as incorrect responses (Adams & Wu, 2002; Martin et al., 2000). In the National Assessment of Educational Progress (NAEP), not-reached items are always ignored (Allen et al., 2001), whereas fractionally correct scores are used for omitted multiple-choice items and omitted non-multiple-choice items are scored in the lowest scoring category (Allen et al., 2001). In the National Educational Panel Study (NEPS), all missing responses are ignored for both item and person parameter calibration (Pohl & Carstensen, 2012). As is evident from these examples, no state-of-the-art procedure for missing data in large-scale assessments exists.

In the last two to three decades, researchers have worked on developing models that are able to account for nonignorable missing responses. In these so called model-based approaches, the probability of a missing value is jointly modeled with the probability of observing a correct response (Glas & Pimentel, 2008; Glynn, Laird, & Rubin, 1986; Heckman, 1976; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Rose, et al., 2010). Lately, the models were

also extended, allowing for further covariates (Glas, Pimentel, & Lamers, in press; Moustaki & Knott, 2000; Rose, 2013). In this way, a person´s *missing propensity* is modeled as an additional manifest or latent variable, and is thus considered in the measurement model for ability. Besides the gain of being able to take nonignorable missing values into account, the models have the feature of being very flexible regarding the estimation of response probabilities (Moustaki & Knott, 2000).

Various studies on missing values in low-stakes assessments evaluated the performance of traditional and recent missing data approaches in the IRT context. These studies are mostly simulation studies. The general idea is to generate matrices with complete data, and invoke missing values in the complete data matrix according to a certain missing mechanism. Afterwards, item and person parameters are estimated for the complete data matrix using the common IRT model and also for the matrices where some values are missing. For evaluating bias, parameters for generating the data are subsequently compared with the estimates retrieved from different missing data approaches. Typically, the amount of missing values and the size of the correlation between the latent ability and the missing propensity is varied. The size of the correlation between the ability and the missing propensity is typically used as an index for the extent of nonignorability in the data (Holman & Glas, 2005; Rose et al., 2010). Many of these simulation studies demonstrated that item and person parameter estimates are biased when missing values are scored as incorrect (Culbertson, 2011; DeAyala, Plake, & Impara, 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010). When ignoring missing values, parameter bias appears only at a certain level of dependence between the missing propensity and the latent ability; at correlations below $r = .4$, the models ignoring missing values usually yield accurate item and person parameter estimates (Glas et al., in press; Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010). Regardless of the

missing data approach, larger amounts of missing values in the data generally result in larger bias.

The previously described studies focused on bias in item and person parameter estimates. In low-stakes large-scale assessment studies, individual competence estimates are typically of less interest since they are not reported back to the examinees. Data users typically aim at answering substantive questions such as competence differences between various groups (e.g., gender, countries) or influences of certain variables (e.g., classroom setting) on competence acquisition. For these analyses, estimates regarding relationships on the group-level are of interest. As previously mentioned, studies found that the missing propensity varies for different subgroups (Köhler et al., in press; Koretz et al., 1993). With regard to comparisons of the general competence level between countries in the PISA 2009 study, for example, data showed that the average missing rate for omitted and not-reached items greatly varied between countries: The average omission rate ranged between 2.6% in the USA and 20.6% in Azerbaijan (OECD, 2012); the amount of not-reached items ranged between 0.3% in the Netherlands and 13.5% in Kyrgyzstan (OECD, 2012). The different missing propensity levels across subgroups can have a major effect on country rankings of the competence level. More importantly, this effect might differ for different missing data approaches. For example, Robitzsch (in press) demonstrated that PISA country rankings change depending on the treatment of missing values. Cosgrove and Cartwright (2014) investigated the large decline of Irish students' mean score in mathematics and reading in the PISA 2009 study. From the amount of missing responses, which drastically increased compared to previous test waves, they concluded that the drop in country ranking resulted from lack of student engagement rather than from lack of student competence. When keeping in mind that for the person parameter estimation in PISA missing values are treated as

incorrect, it is plausible to assume that competencies are underestimated when missing values occurred from lack of motivation rather than from lack of knowledge.

So far, no study exists that systematically investigates under which conditions the different approaches lead to unbiased parameter estimates when analyzing relationships on the group-level, especially when the group-level variable of interest—that is, the explanatory variable—also relates to the missing propensity. The aim of the current study is to evaluate the performance of different approaches to handle missing data with regard to group-level statistics. In this study, we solely focused on the propensity to omit items. We will discuss whether and how the results can be extended to the propensity to not reach items as well. We propose that the magnitude of bias in the group-level statistics varies depending on (a) how the missing values are handled, and on (b) the sizes of the relationships between the ability, the omission propensity, and the explanatory variable. The first proposition is based on results from the previously mentioned studies, which illustrated that different missing data approaches vary in their performance to recover accurate item and person parameter estimates. It is plausible that these differences also emerge on the group-level. When considering the second proposition, the size of the correlation between the ability and the omission propensity should impact the bias, since the higher they correlate the greater the nonignorability of missing values. The size of the correlation between the explanatory variable and the omission propensity might also affect the accuracy of the estimated group-level parameter. Our research is relevant insofar that in large-scale assessments, missing values depend on the ability and on explanatory variables—that is, the variables researchers are interested in and use in their analyses. It is important to evaluate under which missing data conditions which missing data approach leads to accurate outcomes. Our aim was to contribute to establishing an adequate method for dealing with missing values in large-scale assessments, where group-statistics are of major interest.

Simulation Study

Design

The goal of the simulation study was to generate data sets with missing values, where the missing data depended on the ability and the explanatory variable to various degrees. This was done according to the between-item-multidimensional IRT model by Holman and Glas (2005), where the missing data process is modeled as a latent variable: the omission propensity. Therefore, we first generated data sets **X** and **D**, each consisting of $I = 20$ items and $V = 1000$ persons. **X** reflected the latent ability, ξ, and contained the response indicators, x_{iv}, defined as

$$x_{iv} = \begin{cases} 0 & \text{for an incorrect response} \\ 1 & \text{for a correct response} \end{cases},$$

where i indexed the items from $i = 1, \ldots, I$, and v indexed the persons from $v = 1, \ldots, V$. **D** reflected the omission propensity, θ, and comprised the missing data indicators d_{iv}, defined as

$$d_{iv} = \begin{cases} 0 & \text{if } x_{iv} \text{ was omitted} \\ 1 & \text{if } x_{iv} \text{ was observed.} \end{cases}$$

The data sets with missing values **M** resulted from coding each x_{iv} in **X** as missing for each $d_{iv} = 0$ in the respective **D** matrix. The response indicators m_{iv} were thus defined as

$$m_{iv} = \begin{cases} 0 & \text{for an incorrect response} \\ 1 & \text{for a correct response} \\ \text{NA} & \text{for a missing response} \end{cases}.$$

In order to obtain the binary manifest indicators for **X** and **D**, we had to first generate item and person parameters. This was done using the R package mvtnorm (Genz et al., 2014) for the statistical software R (R Development Core Team, 2014). The parameter β_i for the 20 items in **X** represented the difficulty of answering item *i* correctly. The parameter δ_i for 20 items in **D** represented the difficulty of giving a response to the respective item. We drew these parameters from a bivariate normal distribution, so that the difficulties of answering an item correctly and the difficulties of giving a response correlated at r = .5, with means fixed at 0 and variances fixed at 1. For the generation of all data sets, the item parameters were kept constant. The size of the correlation was chosen to be in accordance with the relationship typically found in competence tests of large-scale assessments (see, e.g., Pohl, Haberkorn, Hardt, & Wiegand, 2012; Rose et al., 2010). The person parameters for the ability level, ξ_v, the omission propensity, θ_v, and the explanatory variable, Z_v, were drawn from a multivariate normal distribution. Note that higher values on the omission propensity essentially indicate less omitted items. The correlation between ability and *Z* was fixed at $r(\xi_v, Z_v)$ = .2 for all data sets. Note that this was the coefficient we later aimed to retrieve, and that the amount of variance *Z* explained in ξ was 4% (R^2 = .04). The following parameters were varied: the size of the correlation between the ability and the omission propensity, with $r(\xi_v, \theta_v)$ = 0, .2, .4, and .6, and the size of the correlation between the omission propensity and *Z*, with $r(\theta_v, Z_v)$ = 0, .1, .3, and .5. Note that for the data sets with missing values **M**, which were generated under the condition of $r(\xi_v, \theta_v)$ = 0 and $r(\theta_v, Z_v)$ = 0, the missing data are MCAR. For the combination of $r(\xi_v, \theta_v)$ = 0 and $r(\theta_v, Z_v)$ > 0, the missing data are MAR, and for all combinations where $r(\xi_v, \theta_v)$ > 0, the missing data are MNAR. Keep in mind that as the relationship between ξ_v and θ_v increases, the violation to ignorability of the missing values increases. Regarding the correlation between the ability and the omission propensity, the values were chosen in order to cover a rather wide range of possible correlations. In terms of the

correlation between the omission propensity and *Z*, the conditions of no correlation as well as low, medium, and high sizes of correlations according to Cohen (1988) were covered. A third aspect we varied was the average amount of missing values in the data. We considered a rather extreme case of 50% missing data in order to enhance the effects, and 10% missing data, which more closely maps the typical amount found in actual competence data (see, e.g., Cosgrove & Cartwright, 2014; Koretz et al., 1993; OECD, 2012). Therefore, the means of all three trait variables were fixed at 0 in the 50% missing data conditions, and only the mean of the omission propensity variable was changed to 3 in the 10% condition. The variances of all three trait variables were fixed at 1 in all conditions. Altogether, in generating the data, three factors were varied, resulting in a 4 (correlation between the ability and the omission propensity) x 4 (correlation between the omission propensity and *Z*) x 2 (amount of missing data) design with = 32 cells. The number of replications for each of the possible combinations was $w = 100$, which resulted in 3200 **X** and **D** matrices, respectively. With the generated person parameters ξ_v and θ_v, and the item parameters β_i and δ_i, we calculated the probabilities for a correct response, $p(x_{iv} = 1)$, and the probabilities for giving a response, $p(d_{iv} = 1)$, to the item according to a Rasch model (Rasch, 1960). Note that this was done separately for two sets of data matrices **X** and **D**. We compared the probabilities for a correct response, $p(x_{iv} = 1)$, and the probabilities for giving a response, $p(d_{iv} = 1)$, to values we randomly drew from a uniform distribution on the interval (0, 1). When our obtained probability exceeded the randomly drawn value, the respective response and missing data indicators were scored $x_{iv} = 1$ and $d_{iv} = 1$; otherwise, $x_{iv} = 0$ and $d_{iv} = 0$.

We subsequently used latent regression models to estimate the relationship between the latent person ability variable ξ and the manifest variable Z. We first analyzed the complete data sets **X** using a unidimensional latent regression model in order to establish a reference and to illustrate the relationship between ξ and Z in the complete data. For data sets containing missing

values **M,** we considered three different missing data approaches: (1) including the latent omission propensity θ_v in the measurement model, (2) treating missing values as incorrect responses, and (3) ignoring missing values in the estimation. All 3200 **M** matrices were analyzed applying all three approaches. Note that for the approach of treating missing items as incorrect answers, each m_{iv} = NA was replaced by m_{iv} = 0. The first approach was based on the between-item multidimensional model by Holman and Glas (2005). We included *Z* in form of a latent regression variable. Using marginal maximum likelihood (MML) estimation, which is also the typical estimation method in most large-scale assessments (see, e.g., OECD, 2012; Pohl & Carstensen, 2012; U.S. Department of Education, 1999), the likelihood of the multidimensional model can be expressed as

$$L = \prod_{v=1}^{V} \prod_{i=1}^{I} p(x_{iv}|\xi_v, \beta_i) p(d_{iv}|\theta_v, \delta_i) g(\xi_v, \theta_v|Z_v, \eta, \sum), \quad (3)$$

where $g(\xi_v, \theta_v|Z_v, \eta, \sum)$ represents the density of the common distribution of ξ_v and θ_v, which is expected to be bivariate normal. Z_v is the value of person *v* on the variable *Z*, η are the regression coefficients, and $\sum$ represents the covariance matrix of the residuals. For the second and third approach, unidimensional latent regression IRT models were estimated. Equation 3 thus simplifies to

$$L = \prod_{v=1}^{V} \prod_{i=1}^{I} p(x_{iv}|\xi_v, \beta_i) g(\xi_v|Z_v, \eta, \sigma^2), \quad (4)$$

where σ^2 is the residual variance of ξ_v. All models were estimated using the R package TAM (Kiefer, Robitzsch, & Wu, 2014). For computing the integrals, we used Gauss-Hermite

quadrature with 20 nodes per dimension. A minimum deviance change of .0001 served as the convergence criterion.

In a last step, we calculated the mean standardized regression coefficient across all 100 replications for each of the considered combinations and evaluated R-squared.

Results

The mean standardized regression coefficients from the latent regressions of ability on Z over the 100 replications are depicted in Figure 1. As expected, the mean standardized regression coefficients were close to the generating parameter ($r(\xi_v, Z_v) = .2$) in all conditions for the complete data analyses and also for the analysis where the omission propensity was included in the measurement model. Note that the model including the omission propensity was equivalent to the model generating the data, and was therefore expected to succeed in retrieving unbiased parameter estimates.

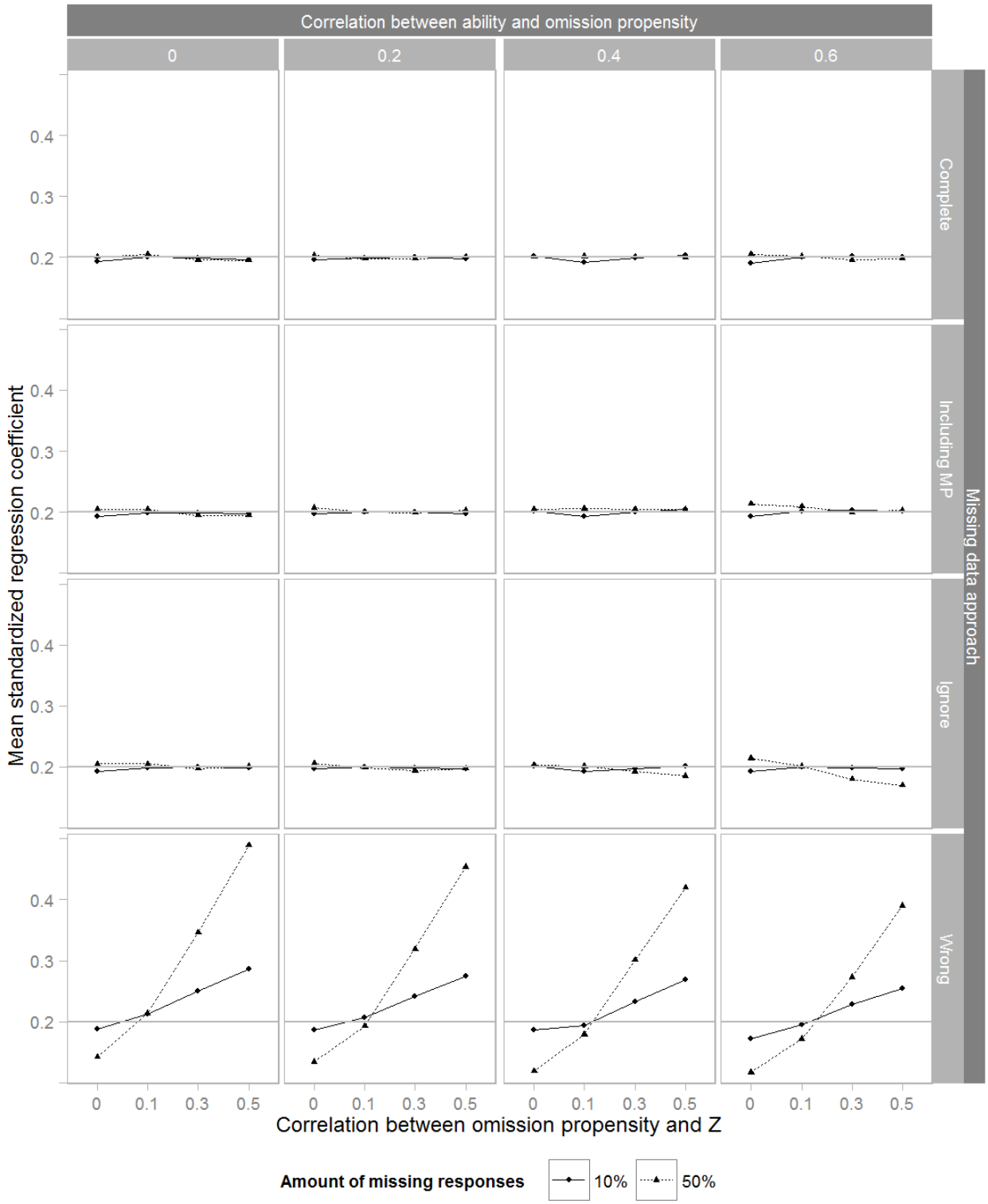

Figure 1. Estimated mean standardized regression coefficients of ability on the variable *Z* across 100 replications for all realized combinations. MP = missing propensity.

Ignoring missing values.

As is evident from the figure, ignoring missing resulted in unbiased estimates of the regression coefficient when the missing values were either MCAR or MAR. Furthermore, the mean standardized regression coefficient was close to the generating parameter in the conditions where the correlation between omission propensity and ability was $r(\xi_v, \theta_v) = 0$ and $r(\xi_v, \theta_v) = .2$. In the conditions where $r(\xi_v, \theta_v) = .4$ and $r(\xi_v, \theta_v) = .6$, the regression coefficients were unbiased for low correlations between the omission propensity and *Z*. For higher correlations between the omission propensity and *Z*, the regression coefficient was slightly underestimated. This underestimation resulted from the fact that the distribution of omitted items in **M** depended on both ability and *Z*, but only *Z* is included in the measurement model.

To illustrate the reason for the underestimation, we analyzed two exemplary data sets (each with 50% missing data) in more detail. The first data set was chosen from the 100 generated data sets where the missing propensity only depended on *Z*, that is $r(\xi_v, \theta_v) = 0$ and $r(\theta_v, Z_v) = .5$ (the mean standardized regression coefficient across all replications was estimated at $b_{std} = .205$, and was thus hardly biased). The second example was chosen from one of the 100 generated data sets where the missing propensity depended on both *Z* and ability, that is $r(\xi_v, \theta_v) = .6$ and $r(\theta_v, Z_v) = .5$ (the mean standardized regression coefficient across all replications was $b_{std} = .170$, and was thus slightly underestimated). The two data sets were chosen so that the estimated standardized regression coefficients in the examples closely resembled the mean standardized regression coefficient over all 100 replications. They were $b_{std} = .199$ for the first example and $b_{std} = .166$ for the second example. We estimated expected a posteriori (EAP; Mislevy & Stocking, 1989) ability estimates for both data sets and evaluated the difference between the true ability parameters and the estimated EAP ability parameters (i.e., bias in parameter estimates) in relation to *Z*. The bias in ability scores in relation to *Z* is depicted in

Figure 2. We first consider the data example where the missing propensity depends on *Z* but not on ability (see Figure 2a): As is evident from the regression line—regressing the difference between the estimated and the true ability score on *Z*—the average bias in ability estimates was close to zero for all levels of *Z*. Therefore, the regression coefficient from the regression of ability on *Z* was unbiased. In the data example where the omission propensity highly depended on both *Z* and ability (see Figure 2b), the average bias was different for different levels of *Z*: The ability estimate was, on average, overestimated for people with lower *Z* scores; they were, on average, underestimated for people with higher *Z* scores. The fact that the average bias in ability estimates was different for different levels of *Z* led to the slight underestimation of the regression coefficient.

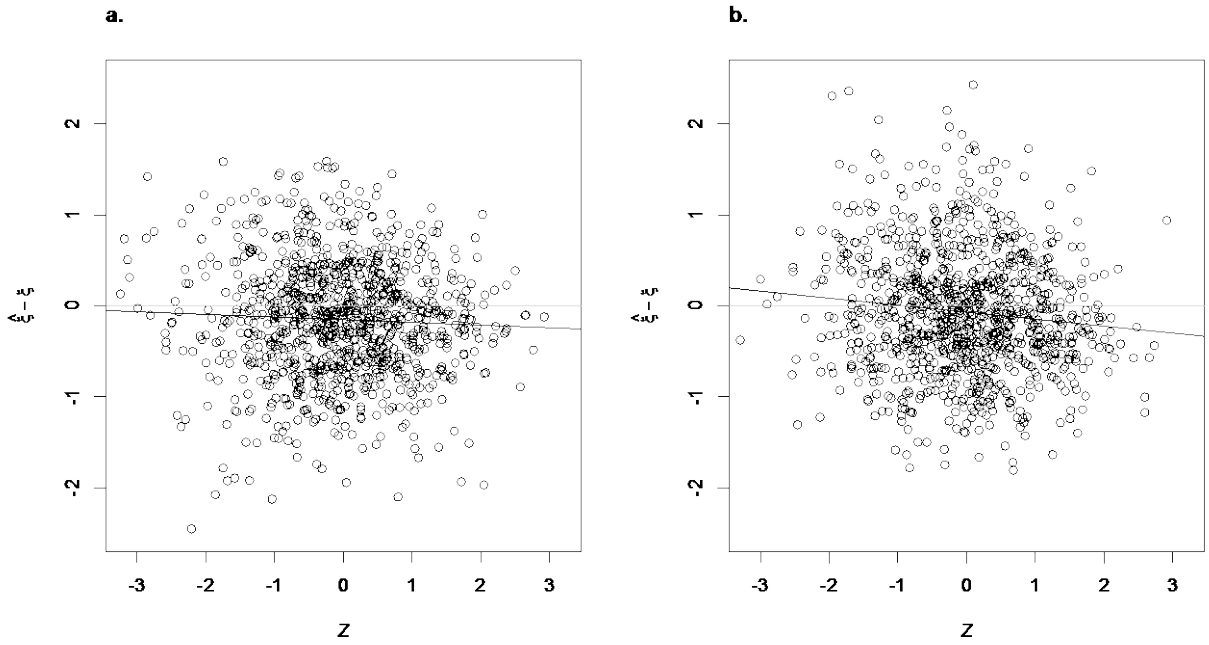

Figure 2. Bias in ability estimates when ignoring missing data for (a) no correlation and (b) correlation of $r(\xi_v,\theta_v)$ = .6 between missing propensity and ability. Correlation between missing propensity and *Z* was $r(\theta_v, Z_v)$ = .5 in both analyses.

When considering the average R-squared across all 100 replications, the conclusion would be that *Z* explained about 3% of the total variance of ability. This differed about 1% from the true explained amount of variance, which was 4%. Ignoring missing values when the missing propensity depended on ability and the explanatory variable had thus a rather small effect on outcomes of group-level analyses. In the case of only 10% missing values, bias was hardly noticeable: Even for the most extreme condition where $r(\xi_v, \theta_v)$ = .6 and $r(\theta_v, Z_v)$ = .5, the estimated explained amount of ability variance was 3.9%, and therefore very close to the true 4%.

Scoring missing values as incorrect.

When treating missing values as incorrect, the estimated regression coefficients were substantially biased. In the conditions with low correlations between the omission propensity and *Z*, the regression coefficients were underestimated. In the conditions where the correlation between the omission propensity and *Z* were greater than .1, the regression coefficients were overestimated (see Figure 1). This occurred regardless of the relationship between ability and missing propensity.

In order to illustrate the cause for the over- and underestimations, we again chose two exemplary data sets (from the conditions with 50% missing values): One data set in which the missing values neither depended on ability nor on *Z*, that is $r(\xi_v, \theta_v)$ = 0 and $r(\theta_v, Z_v)$ = 0 (the mean standardized regression coefficient across all replications was b_{std} = .142, and thus underestimated), the other data set from the condition where the missing values highly depended on *Z* but not on ability, that is $r(\xi_v, \theta_v)$ = 0 and $r(\theta_v, Z_v)$ = .5 (the mean standardized regression coefficient across all replications was b_{std} = .489, and thus highly overestimated). As before, we chose the data sets so that the estimated standardized regression coefficients in the examples resembled the mean standardized regression coefficient across all 100 replications, which was

b_{std} = .142 for the first example and b_{std} = .490 for the second example. We estimated the EAP ability values for each person in each data set and compared them with the true ability parameters. Figure 3 displays the difference (i.e., bias) between the true and the estimated ability values in relation to *Z*. As is evident from the figure, the ability values were, on average, underestimated. This general underestimation resulted from imputing each missing value with a score of 0. Naturally, imputing an incorrect answer for an unobserved correct answer leads to a lower ability score (compared to the ability score if the item had been observed). In our simulation, people with high ability levels, which had more correct answers in the complete data set than people with low ability levels, were more drastically underestimated compared to people with low ability levels. Therefore, people with high ability levels were more drastically underestimated than people with lower ability levels. Furthermore, the average bias in ability estimates was different for different levels of *Z*. In the data example where missing values did not depend on *Z* (Figure 3a), persons had an equal amount of missing values, irrespective of *Z*. However, since the generating parameter for the correlation between ability and *Z* was $r(\xi_v, Z_v)$ = .2, people with higher *Z* scores had higher ability levels, and their ability was therefore, on average, more underestimated than the ability of people with lower *Z* scores. Thus, the standardized regression coefficient when regressing ability on *Z* was underestimated. In the example where missing values highly depended on *Z* (see Figure 3b), ability estimates were, on average, more underestimated for people with low *Z* scores than for people with high *Z* scores, since people with higher *Z* scores obtained substantially less missing values. As a result, the slope of the standardized regression coefficient when regressing ability on *Z* was substantially overestimated.

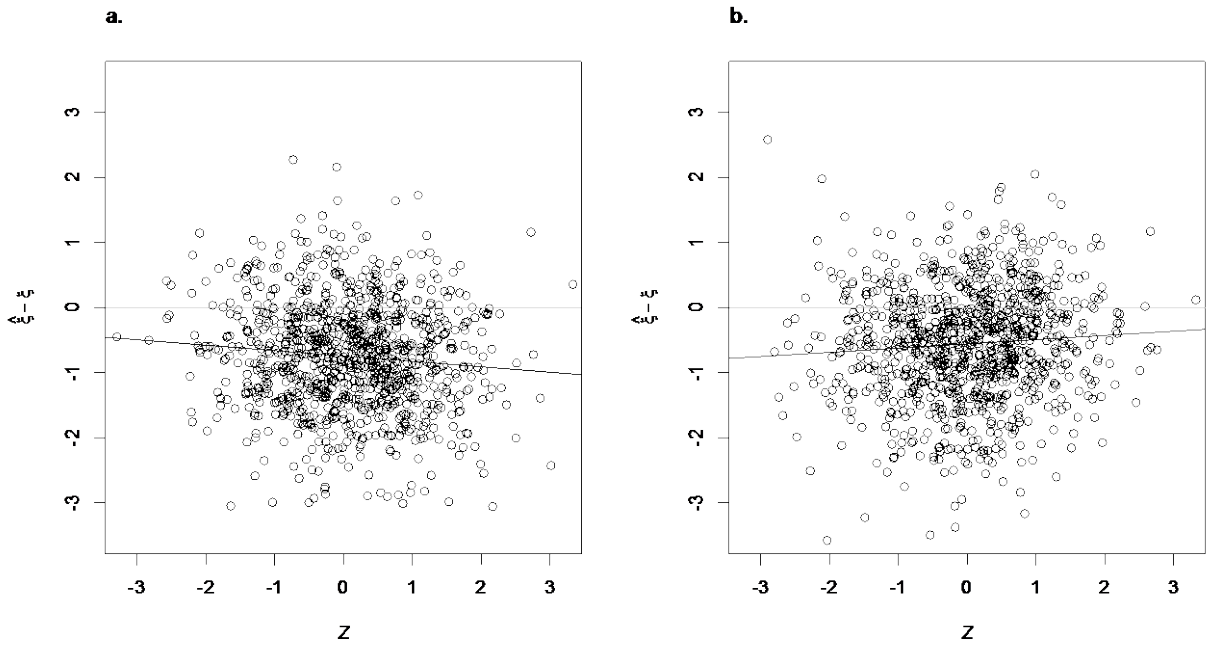

Figure 3. Bias in ability estimates when treating missing values as incorrect responses for (a) no correlation and (b) correlation of $r(\theta_v, Z_v)$ = .5 between missing propensity and *Z*. Correlation between missing propensity and ability was $r(\xi_v,\theta_v)$ = 0 in both analyses.

Considering the mean R-squared across all 100 replications, the impact of treating missing values as incorrect was quite severe. In the condition with 50% missing values and the correlations between ability and missing propensity and between missing propensity and *Z* both set at 0, the conclusion from the regression analysis would be that *Z* approximately explained only 2% of the variance of the ability variable. This was half the amount of the truly explained variance. In the most extreme condition with 50% missing values, $r(\xi_v,\theta_v)$ = .6, and $r(\theta_v, Z_v)$ = .5, the estimated explained variance was 24%, which was a large overestimation of the actual 4%.

The size of the correlation between the omission propensity and the ability also influenced the estimated regression coefficient when treating missing values as incorrect. For a constant correlation between the omission propensity and *Z*, the estimated regression coefficient decreased

as the correlation between ability and the omission propensity increased. This was mainly due to the fact that the estimated total variance of ability was greater in the conditions where ability depended on *Z* as well as on the omission propensity. The proportion of variance *Z* explained in ability was therefore comparably smaller, thus reducing the standardized regression coefficient.

In sum, the results indicate that for data where the missing process is similar as in our simulated data sets, bias in parameters of group-level analyses is systematic and quite large when treating missing values as incorrect answers. When ignoring missing values, bias is more severe for high correlations between ability, the missing propensity, and *Z*, but still produces quite acceptable estimates. The approach of including the omission propensity in the measurement model retrieves accurate parameters.

Empirical Example

Method

Finally, we applied the different missing data approaches to empirical data in order to elucidate whether our results were transferable to real competence test data, and to examine how the coefficients from group analyses are affected when the missing values relate to ability and the explanatory variable. We chose the mathematics domain in the adult cohort of the NEPS data (Blossfeld, Roßbach, & von Maurice, 2011), and investigated the relationship between gender and ability in mathematics when using different missing data approaches. We chose gender as the group-level variable, since preliminary analyses showed that both omission propensity and ability relate to gender (Köhler et al., in press). The number of items in the mathematical competence test was 21, the number of examinees 5,298. As we only focused on omitted items in our simulation study, we excluded examinees with not-reached items from the analysis, resulting in 2,333 persons. We applied the three different missing data approaches—(1) including the omission propensity in the measurement model, (2) ignoring missing values, and (3) treating

missing values as incorrect responses—to estimate the standardized regression coefficient of mathematical ability on gender. The average amount of omitted items was only 4%. We therefore expected similar but weaker effects in the real data example as compared to our simulation study.

Results

In our data example, the correlation between the difficulty of answering an item correctly and the difficulty of giving a response to the respective item was $r(\beta_i, \delta_i) = .67$, meaning that the probability for giving a response highly depended on the difficulty of the item. The correlation between ability and omission propensity was substantial with $r(\xi_v, \theta_v) = .49$; the missing values were therefore nonignorable to a considerable amount. The correlation between the omission propensity and gender (0 = female; 1 = male) was $r(\theta_v, Z_v) = .31$, meaning that women omitted substantially more items than men. Note that the size and direction of these correlations resemble the condition in the simulation study were omitted values were MNAR and depended on both the ability and the explanatory variable. We therefore expected similar outcomes with regard to the regression coefficient when applying different missing data approaches. The estimated standardized regression coefficients were (1) $b_{std} = .329$ ($b_{unstd} = 0.893$, $SE = .041$) when including the omission propensity, (2) $b_{std} = .327$ ($b_{unstd} = 0.886$, $SE = .042$) when ignoring omitted values, and (3) $b_{std} = .339$ ($b_{unstd} = 0.947$, $SE = .043$) when treating omitted values as incorrect responses. This illustrates that the estimated mean ability difference between male and female examinees slightly deviates depending on the applied missing data approach. The difference was small between the first two approaches, and not statistically significant. The regression coefficients from the first two approaches significantly differ from the regression coefficient when treating omissions as incorrect responses.

The results mirror the findings in our simulation study, where, in the more extreme conditions, the approach of ignoring missing values resulted in slightly lower regression

estimates, and the approach of treating missing values as incorrect responses in higher estimates than in the approach where the omission propensity was included. The differences were rather small. Keep in mind, however, that the amount of omitted items was low. The effects would probably be more severe in competence tests with higher omission rates.

Discussion

The goal of the current study was to assess the performance of various missing data approaches in appropriately recovering parameter estimates of group level-analyses. It was of particular interest to evaluate how the approaches perform when the missing data are either MCAR, MAR, or MNAR, and relate to the explanatory variable. We varied the amount to which the missing values depended on ability and the explanatory variable in order to evaluate under which conditions the various approaches still manage to retrieve accurate parameter estimates. We applied two methods commonly employed in large-scale assessments, that is ignoring missing values and treating them as incorrect, as well as a recently developed model-based approach that deals with data that are MNAR.

The three approaches differ in their performance, depending on the missing data mechanism present in the data. The model-based approach accurately estimates the group-level relationship regardless of the missing data process. This approach thus accurately accounts for missing data that is MAR as well as for missing data that is MNAR. The approach of ignoring missing values is appropriate when the amount of missing values in the data is small. For large amounts of missing values, the approach also yields unbiased results when the MAR assumption is not largely violated, that is, when the probability for a missing value does not greatly depend on the latent ability variable. Treating missing values as incorrect leads to considerable bias in parameter estimates of group-level analyses, either over- or underestimating the relationship between ability and the explanatory variable. In the case where the missing data process in large-

scale studies is similar to the one we induced in our simulation, treating missing values as incorrect should not be employed.

Our results confirm and enhance previous research. In line with findings regarding bias on person and item parameters, the approach of ignoring missing values leads to accurate estimates of group-level parameters when the correlation between the missing propensity and the ability is low. For correlations above $r = .4$, researchers working with competence test data should consider a model which takes the missing propensity into account. This is especially relevant when missing values also depend on the explanatory variable. With regard to treating missing values as incorrect responses, we showed that results found for person parameters (Culbertson, 2011; DeAyala et al., 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Holman & Glas, 2005; Pohl et al., 2014; Rose et al., 2010) are also present to a considerable amount on the group-level. If, for example, two countries truly differ in their ability levels, and the country with the lower mean ability level also has higher missing rates due to lack of motivation, the average ability difference between the countries will probably be overestimated. In our analyses we found that ignoring missing values leads to considerably less bias.

For high-stakes assessments, ignoring missing values should not be the method of choice, since examinees aware of the scoring method mighty simply omit the questions they are unsure of. This would drastically increase the missing data rates. In low-stakes assessments, however, examinees typically do not aim at increasing their test score. Our results also demonstrate that even when the missing propensity highly depends on the ability, meaning that test takers mostly omit items they would have answered incorrectly, the ability parameters remain relatively unbiased. For very high correlations between missing propensity and ability, researchers should rather consider including the missing propensity in the scaling model than treating missing values as incorrect.

There are some limitations to our study and the inferences that can be drawn from it. In all simulation studies, data are generated according to certain models. Of course, missing values could be simulated according to a different mechanism. An alternative would be, for example, to induce a missing value depending on whether the response to the item was correct or incorrect (Robitzsch, in press). Different data generating models may result in different conclusions about the missing data approach. Since studies found that some of the reasons for missing values include inability to answer the item correctly as well as other personality states and traits (Jakwerth, Stancavage, & Reed, 1999; Köhler et al., in press; Koretz, et al., 1993; Pohl et al., 2014; Rose et al., 2010), we generated the missing data depending on ability and an explanatory variable, examining whether a disregard of these relationships affects group-level parameter estimates. A simulation study cannot mirror all properties of the missing data process in actual empirical data. We tried to cover various missing data scenarios and missing data mechanisms, in order to depict possible scenarios in empirical data. Furthermore, we used an empirical example to validate our conclusions.

We only considered the case where the missing data depend on the ability and an explanatory variable. In empirical data, however, the missing propensity may depend on several other person characteristics (Jakwerth et al., 1999; Köhler et al., in press; Koretz et al., 1993). It would be interesting to assess how the group-level estimate is affected when these relationships are accounted for in the data generation and analysis. Furthermore, we only considered missing values due to omitted items. The propensity to not reach items typically differs from the propensity to omit items, and should be handled separately in the scaling (Moustaki & O'Muircheartaigh, 2000; Pohl et al., 2014; Rose, 2013). It is worthwhile to consider not-reached and omitted items simultaneously and to investigate how the different missing data processes influence parameter estimates from group-level analyses. Other aspects not examined in the

present study are polytomous items or items with different response formats. Though our results probably generalize to polytomous items, the response format might be an issue with regard to the missing propensity. Note that the missing values in our simulated data were induced according to the unidimensional Rasch model (Rasch, 1960). In the literature, there is some indication that persons' tendencies to omit items might be multidimensional with respect to different item response formats (Köhler et al., 2014; Pohl et al., 2014). Future studies might pursue the question of whether and how multidimensionality of the omission propensity affects group-level parameter estimates.

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report.* Paris, France: Organisation for Economic Co-operation and Development.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES Publication No. 2001-509). Washington, DC: National Center for Education Statistics.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (eds.) (2011). Education as a lifelong process – the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Special Issue 14*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cosgrove, J., & Cartwright, F. (2014) Changes in achievement on PISA: the case of Ireland and implications for international assessment practice. Large-scale Assessments in Education, 2, 1-17. doi:10.1186/2196-0739-2-2

Culbertson, M. (2011). *Is it wrong? Handling missing responses in IRT.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

DeAyala, 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,* 213-234. doi:10.1111/j.1745-3984.2001.tb01124.x

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225-245. doi:10.1111/j.1745-3984.2008.00062.x

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2014). mvtnorm: Multivariate normal and t distributions. R package version 1.0-2. Retrieved from http://CRAN.R-project.org/package=mvtnorm

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68,* 907-922. doi:10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (in press). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling.*

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples*. New York, NJ: Springer. doi: 10.1007/978-1-4612-4976-4_10

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, limited dependent variables, and simple estimator for such models. *Annals of Economic and Social Measurement, 5,* 475-492.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71,* 732-746. doi:10.1177/0013164410390032

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,* 1-17. doi:10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions* (NAEP Validity Studies, Working Paper Series). Palo Alto, CA: American Institutes for Research. Retrieved from http://www.air.org/sites/default/files/downloads/report/Jakwerth_report_0.pdf

Kiefer, T., Robitzsch, A., & Wu, M. (2014). TAM: Test analysis modules. R package version 1.5-2. Retrieved from http://cran.r-project.org/web/packages/TAM/index

Köhler, C., Pohl, S., & Carstensen, C. H. (2014). Taking the missing propensity into account when estimating competence scores—Evaluation of IRT models for non-ignorable omissions. *Educational and Psychological Measurement*. doi: 10.1177/0013164414561785

Köhler, C., Pohl, S., & Carstensen, C. H. (in press). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling.*

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Tech. Rep. No. 357). Los Angeles, CA: Center for Research on Evaluation, Standards and Student Testing.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College.

Mislevy, R. J., & Stocking, M. L. (1989) A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75. doi: 10.1177/014662168901300106

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

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: Statistics in Society, 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*, 259-276.

OECD (2012). *PISA 2009 Technical Report, PISA*, OECD Publishing.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 162*, 177-194. doi: 10.1111/1467-985X.00129

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report: Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in Item Response Theory models. *Educational and Psychological Measurement, 74*, 423–452. doi: 10.1177/0013164413504926

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Nielsen & Lydiche. (Expanded edition, 1980)

Robitzsch, A. (in press). Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment. Preprint retrieved from https://sites.google.com/site/alexanderrobitzsch/publikationen

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report ETS RR-10-11). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592. doi:10.1093/biomet/63.3.581

U.S. Department of Education (1999). *The NAEP guide* (NCES Report No. 2000-456). N. Horkay (Ed.). Washington, DC: National Center for Education Statistics.

Chapter 5

Discussion

5 Discussion

The present dissertation addresses the incompatibility between current missing data practices in competence tests of large-scale assessment studies and theoretical considerations on adequate missing data treatments. Model-based approaches that have been proposed more recently might bridge this gap. We investigated these approaches in light of their applicability and effectiveness in the context of low-stakes large-scale assessment studies. Three separate research studies were conducted, which coherently answered the two main objectives of this dissertation. The first aim was to learn about the properties and features of the missing data mechanism in competence tests of large-scale assessment studies. The second focus lay on evaluating model-based approaches as well as the commonly employed approaches in terms of reliable competence measurement.

The first section of this chapter (5.1) summarizes the main results from the three research studies, puts them in context of the overall research topics, and answers the research questions that were raised in the first Chapter (see Section 1.4). These findings are subsequently discussed in light of educational measurement, with an emphasis on implications for the scaling of low-stakes large-scale assessment studies (5.2). The section points out potential advantages and risks of various missing data approaches, and weighs the different options. Lastly, future research perspectives which might aid in evaluating the alternative missing data methods are presented (5.3). Also, further advancements in educational measurement and the resulting new research prospects with regard to missing data are discussed.

5.1 Summary of Main Findings

In this section, the main conclusions are summarized collectively, taking a holistic view of the findings from the three studies. I derived four central points from our results, which are consecutively presented and discussed with regard to the formerly proposed research questions.

The first three points revolve around the assumptions and the features of the missing data process; the last point focusses on the question regarding the necessity of model-based approaches in large-scale assessments.

1. The assumption that the latent omission propensity and the latent ability are bivariate normally distributed is violated. Models with more lenient distributional assumptions better describe the data, and thus more appropriately capture the joint latent skill space. (Study 1)

These findings indicate that, in order to appropriately account for missing responses due to omissions, models that allow for more flexible distribution assumptions should be employed. This can be accomplished by using models assuming discrete latent variables, such as latent class models and general diagnostic models (GDM; von Davier, 2005a). Since only the distribution of the omission propensity is skewed, a latent class model for the omission propensity could interplay with a latent trait model for ability (see Katsikatsou, Kuha, & Moustaki, 2014).

2. The response format has a major impact on the omission process. The mechanism to omit items with simple multiple-choice format is quite distinct from the mechanism to omit items with more complex formats such as complex multiple-choice items (i.e., items with several subtasks, each containing two response options) and matching task items (i.e., several statements need to be matched with corresponding figures or paragraphs). The respective omission processes share some common variance, but differently relate to other stable person characteristics. (Studies 1 and 2)

The fact that the occurrence of omissions depends on the response format of the item is relevant with regard to modeling the missing data mechanism. In the proposed latent model-based approaches, the omission propensity was typically modeled as a unidimensional latent variable (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999). If unidimensionality is

falsely assumed, the model-based approach fails to correctly account for nonignorable nonresponse (Glas et al., in press; Rose, 2013). In order to appropriately include the missing data mechanism in the measurement model, a multidimensional model for the omission propensity can be utilized (see, e.g., Glas et al., in press; Rose, 2013), where the different dimensions reflect the omission propensities on various response formats. Such a multidimensional model captures the missing data process for omitted items more accurately. Furthermore, it adequately takes the relationships between omission propensities, ability, and other variables in the measurement model into account.

3. The propensities to omit and not reach items can be considered relatively stable person characteristics. They are highly related across different competence domains, and also relate to further stable person characteristics such as competencies and demographic variables. The same relationships were found across several age cohorts. The missing propensities thus largely depend on person-specific characteristics. (Study 2)

The stability of the missing propensities confirms the person-specific aspects of the missing propensities. According to these results, it might suffice to only model the missing propensities once when including them in the scaling (compared to modeling the missing propensities separately for each competence domain in the measurement model). The related person characteristics demonstrate which variables are relevant in the model in order to adequately account for missing values. Our findings also benefit simulation studies. When generating data sets with missing values, the missing data need to be induced according to a certain mechanism. The fact that several variables influence the probability for a missing response should be considered and possibly included in the process of generating missing values. Simulation studies which closely map the missing process of empirical data allow for generalizable conclusions on the performance of different missing data approaches.

4. Various missing data approaches differently bias individual ability estimates and group-level parameter estimates.
    a. For more lenient distribution assumptions, individual parameter estimates from models including the omission propensity showed slight deviations to estimates from models simply ignoring the missing data. (Study 1)
    b. The multidimensionality of the propensity to omit items hardly affected individual ability parameters. Including the omission propensity unidimensionally in the measurement model resulted in similar ability parameter estimates as models including the omission propensity multidimensionally. (Study 1)
    c. On the group-level, models including the missing propensity and models ignoring the missing propensity lead to almost identical results, even when the condition of ignorability was strongly violated.[8] (Study 3)
    d. Models where missing values were treated as incorrect responses greatly biased parameter estimates on the group-level. (Study 3)

The results on different individual parameter estimates when comparing models ignoring missing data with models including the omission propensity suggest that information from the missing data mechanism is relevant for the scaling of competencies, and should be considered (see 4a). The dimensionality of the omission propensity seems hardly relevant when modeling the missing data mechanism for omitted items (see 4b). The results on the group-level indicate that the model simply ignoring missing data might suffice for the scaling (see 4c). Treating missing values as incorrect responses inadequately represents group-level relationships, and cannot be

[8]Note that in the study on group-level parameters, we assumed normal distributions for unidimensional models and bivariate normal distributions for multidimensional models. The missing data were also induced according to the model assuming a bivariate normal distribution.

considered a proper approach for dealing with missing data (see 4d). Note that points 4a and 4c seem in conflict with each other to some extent. Keep in mind, however, that bias on the individual level was only minor, and that the scaling models in the two studies differed slightly (see Footnote 6). Also, the group-level parameter might not drastically reflect bias on the individual level, since individual bias can even out if some examinees are over- and others are underestimated. Overall, ability scores are less significant in low-stakes large-scale assessments compared to group scores, so the emphasis should lie on the results from the group-level analyses. Implications and recommendations for the scaling of competence tests are discussed in the next section.

5.2 Implications for Research and Application in Educational Measurement

This section draws inferences from the previously outlined main results. It discusses different missing data approaches and their consequences when applying them in studies on educational measurement. It first focusses on the common missing data approaches of ignoring missing values and treating them as incorrect (5.2.1). The main emphasis lies on ignoring missing values, since this method is hardly implemented in the scaling of competencies, although it leads to more accurate parameter estimates than treating missing values as incorrect. Subsequently, I discuss model-based approaches as a superior method, and debate possible ways of implementing them (5.2.2). These considerations include strengths and limitations with regard to the scaling of competence tests in large-scale assessments (5.2.2). Lastly, I provide further alternatives that allow enhancing measurement accuracy (5.2.3).

5.2.1 Applying common missing data approaches to large-scale assessment data

The research results from the present dissertation aid in determining proper missing data treatments in large-scale assessment studies. Several aspects concerning their implementation in large-scale assessments need to be considered, which are discussed in the following. The main

priority should lie in providing accurate parameter estimates, thus ensuring adequate inferences from data analyses. So far, most studies use the method of incorrect scoring for the scaling of competencies, although criticism of this approach has been voiced early on (Lord, 1974). In accordance with past research (see Culbertson, 2011; de Ayala et al., 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Pohl et al., 2014; Rose et al., 2010), results from the studies conducted within the scope of the present dissertation show that treating missing values as incorrect responses greatly biases parameter estimates, whereas ignoring them has hardly any negative effect. This raises the question of why large-scale assessment studies refrain from implementing this form of missing data treatment. One major aspect to consider is that test assessment and the scaling thereof cannot be viewed in isolation from examinee behavior: Examinees might adapt their test-taking behavior according to how missing values are treated, especially if their motivation to perform well on the test is high. The studies of the present dissertation were conducted using data from a low-stakes assessment study, in which examinees were most likely unaware of the scaling method or indifferent to it, since it had no particular effect on their individual academic careers. For assessments that have major impacts on educational decisions and are relevant for the participating countries, however, the interest to perform well might be more enhanced. In several countries, schools actually prepare their students for the PISA assessment, showing them exemplary item tasks and giving them instructions on test-taking strategies (Department of Education, Early Learning and Culture, 2012; Stephen, 2013). Thus, although PISA is actually labeled low-stakes, it might be considered a high-stakes assessment due to its major global influence. If schools become aware that missing values are ignored in the scaling, they might advise students to only respond to questions they are confident in answering correctly, thus using the scaling method to their advantage. The fear of

many researchers is that if examinees skip all items they would have answered incorrectly, their test scores might become inflated (Culbertson, 2011; Lord, 1974).

So far, the consequences of this type of test-taking behavior have not been thoroughly investigated or discussed. Rose et al. (2010) showed that ability estimates from simulated respondents in the lower and middle ranges of the ability spectrum are slightly overestimated when they omit items they do not know the correct answer to. However, these results were derived from simulated data with almost 50% missing values, and missing propensity and ability correlated very highly (r = .8). In analyses on data sets with fewer missing responses (30%) and a lower correlation between the proportion of correct answers and the observed items responses (r = .6), the model ignoring missing data yielded acceptable results. These findings are in accordance with the results from the third study of the present dissertation, where even for high correlations and high missing data proportions, the models ignoring missing data retrieved accurate parameter estimates on the group-level.

In light of these results, the fear of abuse of the scaling method seems, to some extent, unreasonable. Even if examinees mainly omitted items they feel insecure about, models ignoring missing values in the scaling would be quite robust to this violation of ignorability. However, the previously described simulation studies might not have adequately mapped the true underlying missing data mechanism that would exist in actual data if examinees tried to profit from the scaling method. Skipping items they are unsure of or leaving several items at the end of the test unanswered also gives examinees more time to work on other items, thus increasing the probability for a correct answer on those items. Furthermore, applying this form of test-taking strategy might depend on other cognitive capabilities of the person, such as the ability to evaluate the probability of success on a respective item. Students less capable of self-evaluation might end up skipping more items than they should—in case they actually know the correct answer but are

very insecure about its correctness, and thus refrain from answering it (see also Jakwerth et al., 1999). Moreover, schools might differently prepare their students with regard to the scaling method, which would result in advantages of some schools over others.

In addition to concerns regarding unbalanced advantages when ignoring missing values, the method might also lead to an overall increase in missing responses. Although simulation studies showed that the approach retrieves accurate parameter estimates even for high amounts of missing values in the data, the reliability of the person parameter estimates decreases as the amount of missing values increases (Rose, 2013; Rose et al., 2010). When different people leave different items unanswered, they basically create their own subtest, which consists of only the items they responded to (Rose, 2013). Consequently, the variance in the observed answering patterns is quite large. Due to the vast number of possible observed subtests, relatively few examinees respond to the exact same items. Since each ability estimate and the corresponding standard error for an examinee is based on the items and the examinees of the respective subtest, the standard error for each ability estimate, and hence the overall error variance of the latent ability, is increased. For the sake of highly reliable person parameter estimates, it is therefore desirable to keep the amount of missing values as low as possible (Rose, 2013).

Given the case that examinees aim at maximizing their test score and they manage to accurately self-assess their ability to correctly answer an item, the amount of incorrect scores would drastically decrease. This effect would result in computational problems when estimating the parameters of the IRT model. If examinees only produced correct scores and missing values but no incorrect scores, there would be hardly any variation in the data. In this event, no distinction between more or less difficult items or between more or less skilled people is possible, and the item and person parameters are inestimable. A successful realization of this type of test-

taking behavior would thus certainly threaten the validity and reliability of the measurement model.

Up to this point, the consequences of changing the scaling method from incorrect scoring to ignoring missing values are difficult to predict. Would test takers indeed change their strategy? Would the amount of incorrect scores drastically decrease, or are examinees rather inapt at predicting their probability of success? How severely would the amount of missing data increase? These questions remain unanswered. What can be said for certain thus far is that the method of incorrect scoring should be reconsidered and, ideally, replaced. The approach of ignoring missing values will possibly result in a different test-taking behavior, thus altering the missing data mechanism.

5.2.2 Implementing model-based approaches in large-scale assessments

The option of treating missing values as incorrect responses seems infeasible in light of the negative consequences for parameter estimation. The option of ignoring missing values is impracticable due to the possible negative effects with regard to examinee behavior if the scaling method becomes known. The following subsection thus discusses the option of applying model-based approaches to cognitive data of large-scale assessments. When applying this method, it would be difficult for schools to advise their students on how to enhance their test scores, since model-based approaches detect the missing data mechanism and take it into account. They are thus rather unsusceptible to manipulation. This leaves the question on the adequate model-based approach with regard to real data application. To answer this question, the first three of the main conclusions that pertain to the features and characteristics of the missing propensity (see Section 5.1) will be thoroughly discussed.

First, the results on different distribution assumptions suggested that models with more flexible assumptions should be applied. However, this limits the scope of potential software

programs, since only few of them allow for discrete or mixed distributions (see, e.g., *mirt*, Haberman, 2013; *mdltm*, von Davier, 2005b; *WINMIRA*, von Davier, 2001). Also, some desirable features such as item-fit indices or the estimation of certain person parameter estimates (e.g., plausible values) are not always available in these programs. This makes the model-based approach less convenient. One option would be to disregard the violation to the bivariate normal distribution assumption, although researchers should keep in mind that this slightly affects person ability estimates.

Second, the different omission mechanisms on items with various response formats should be considered. The results from the first study of the present dissertation showed only minor discrepancies between ability estimates from models including a unidimensional or a two-dimensional omission propensity. Therefore, models containing only one dimension for the omission propensity might suffice, making the estimation less computationally extensive. It should be noted, however, that NEPS data, which we used in our study, only comprise few items that are not simple multiple-choice. Studies with many open response items or items with complex response formats might consider a two-dimensional approach for including the omission propensity.

Third, the relative stability of the missing propensities to omit and not reach items suggests that it might suffice to include the respective propensities only once in the overall measurement model. This raises the question of which competence domain should be used for modeling persons' missing propensities. Domains containing more missing values provide more information on examinees' missing propensities, and might be preferable over domains with few missing values. However, augmenting ability scores in domains with hardly any missing data is also questionable. If one domain has many missing values and others do not, an inclusion of the missing propensities from the domain with many missing responses might bias the inferences on

ability in the other domains. Whether the missing data mechanism from a particular domain actually well represents the missing data mechanisms in other domains should be tested before including it in the overall measurement model. Our approach (see Chapter 3) serves as a convenient method for checking this assumption.

Furthermore, variables influencing the missing data mechanism should be considered in the measurement model. Since most large-scale assessment studies include background variables in the scaling model for ability estimation (e.g., plausible values), these variables are already included and can thus simply serve as explaining variables for the missing propensity as well. Note that some studies might not assess all variables that are relevant predictors for the missing propensities to omit and not reach items. Neglecting these in the analysis might slightly affect ability estimation (Glas et al., in press).

Overall, the computational complexity and the gain in accuracy of parameter estimates need to be weighted. The advantage of model-based approaches over common methods most likely remains even if not all features and characteristics of the missing mechanism are considered. A simple unidimensional model for the omitted data and a manifest measure of the not-reached propensity would be easy to implement in the measurement model. Disregarding some aspects and features of the missing data mechanism (such as violations to the distributional assumption, multidimensionality of the missing propensity, some differences between missing mechanisms in different competence domains, and some relevant explaining variables) diminishes the precision of the missing data model. Even if the full scope of the actual missing data process is not considered, however, simple models also take the dependency between the unobserved nonresponse and the unobserved ability into account. Modeling the joint distribution of missing propensity and ability can be considered the main virtue with regard to nonignorable missing values. Although this distribution could be modeled more precisely using more elaborate

models, the general dependency between the variables is taken into account in the less complex model as well. It might thus be considered a compromise between precision and parsimony of the model. Altogether, model-based approaches—including the simpler but less precise models—represent a superior alternative to the common missing data methods.

5.2.3 Further options for dealing with missing values in large-scale assessment studies

In order to bypass the dilemma of choosing between different scaling methods, another option for providing ability measures from large-scale assessments involves estimating parameters with different missing data approaches. The PISA country rankings, for example, could be established using the method of ignoring missing data, the approach of treating missing values as incorrect, and a model-based approach. Subsequent comparisons between the country rankings inform about the influence of the different missing data treatments on the ability estimates. This would draw attention to the necessity of considering missing values and make the uncertainty of measurement due to missing values rather transparent. Supplying several rankings might, however, appear rather ambiguous to the public and also complicates interpretation of the results. Instead of providing country rankings for each missing data method, another approach could be to decide on one specific missing data method, and report estimation error due to missing values. Just like sampling error and measurement error, missing values can be considered a source of uncertainty in the measurement. When establishing PISA country rankings, for example, upper and lower ranks for each participating country are estimated, taking account of measurement and sampling error. These boundaries could be extended so they comprise estimation error due to missing responses. Thus, the results can be interpreted with due caution.

So far, the present dissertation focused on methodological approaches for the treatment of missing data in large-scale assessment studies. Several options outside the scaling exist in order to reduce the missing data problem. One primary objective of test developers is to increase

examinee response (see, e.g., Jakwerth et al., 1999). This should be considered when outlining the design of the study. Possible options to improve response rates include (a) encouraging examinees in the instructions of the study to respond to all items (see, e.g., Culbertson, 2011), (b) giving them incentives for participating in the study, hence increasing their motivation (see, e.g., (Baumert & Demmrich, 2001), and (c) forcing them to answer each question. The latter is hardly realizable in paper-and-pencil tests, but can be implemented in computer based assessments. This is accomplished by allowing examinees to click the *next* button only if they gave a response to the item at hand.

With regard to encouragement (a), most study manuals already contain instructions that advise students to thoroughly work on the test items. Students are also recommended to provide an answer even if they are unsure about its accuracy (PISA Project Consortium, 2003). However, large amounts of missing values still exist, which possibly cannot be prevented by any form of instruction. It would be worth to consider more straightforward instructions, however. Examinees could be informed that incorrect answers do not lead to a deduction of points. Also, they should be aware that leaving items unanswered leads to incorrect scoring, and that they are working against their best interest if they refrain from guessing. In terms of incentives (b), few studies were conducted in the area of low-stakes assessments. They showed that rewards actually hardly affect student performance (Baumert & Demmrich, 2001; Berlin et al., 1992; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005). Though incentives generally enhance compliance to participate in surveys (Berlin et al., 1992; Church, 1993; Martin, Helmschrott, & Rammstedt, 2014), they hardly impact item response rates (Berlin et al., 1992). After people decide on taking part in a study—for whatever reason—they seem equally motivated to participate. Forcing people to respond (c) deters them from omitting items. This method has potential drawbacks, however. Examinees might, for example, become frustrated not being able to switch back and forth

between the items. Also, if they have no idea as to how to respond to an item with an open ended response format, they might stop responding altogether, terminating the test. This potentially increases the drop-out rate, resulting in larger amounts of not-reached items. All in all, nonparametric approaches seem rather limited in terms of minimizing the occurrence of missing values. Nevertheless, these options should be fully exhausted, since a reduction of missing values increases measurement reliability.

One further parametric approach which received quite little attention is Lord's (1974) suggestion to impute an omitted item with a correct score at chance level. This method has not been applied in practice, although it yielded quite promising results (Culbertson, 2011; de Ayala et al., 2001). De Ayala et al. introduced a slightly amplified method: Instead of calculating the reciprocal of the number of response options, they replaced the probability of a correct response with .5. Thus, the probability for a wrong prediction (e.g., a correct response is imputed although the examinee would have answered the item incorrectly) is reduced to .5. De Ayala et al. received slightly better results with their method than with the reciprocal of the number of response options. Culbertson (2011) tried to replicate these findings, but only received unbiased parameter estimates if the omissions were unrelated to the actual response. Similar to the approach of ignoring missing values, people can artificially inflate their score by only responding to items they are confident of answering correctly. Keep in mind that imputation methods fail at taking the dependency between missing values and ability into account. Therefore, this approach seems also rather limited. Using the reciprocal of the number of response options, however, might not influence the behavior of test takers drastically—compared to ignoring missing values and using .5 as chance level—since the probability to guess correctly equals chance level. Examinees thus would not profit from leaving items unanswered. In fact, their probability to guess correctly might exceed chance level if they can identify some of the distractors. Lord's (1974) proposition

might thus still be considered a reasonable alternative approach to handle omitted items, and could easily be implemented in high-stakes as well as low-stakes assessment studies. Such a model could also include item and examinee characteristics (Lord, 1983; Mislevy & Wu, 1996). Putting this method into perspective with other missing data approaches, I consider it inferior to the model-based approach with regard to accuracy of parameter estimates, since it fails to take nonignorable nonresponses into account. However, no missing data mechanism needs to be specified, and the method is therefore more convenient. It might be considered superior to treating missing values as incorrect, since it results in more accurate parameter estimates (de Ayala et al., 2001; Culbertson, 2011), and possibly also to ignoring missing values, since it is less susceptible to abuse.

In light of these imputation approaches, newer imputation methods should also be mentioned. These include multiple imputation (Rubin, 1987), two-way imputation (Bernaards & Sijtsma, 2000), response-function imputation (Sijtsma & van der Ark, 2003), conditional mean imputation (Schafer & Schenker, 2000), and the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). These approaches yield reasonable results, although they fail to account for nonignorable nonresponse as well (Rose, 2013). The methods are rather complex to apply, though some of them are already implemented in statistical software programs (e.g., SAS, SPSS). Thus, the data user might profit from this approach, since it avoids specifying a model for the missing data process. As already discussed in the previous paragraph, the method might be considered superior to ignoring missing values and treating them as incorrect. In comparison with model-based approaches, they might lack accuracy when missing data are nonignorable.

5.3 Research Perspectives

The current section discusses present as well as future research perspectives on missing data treatment. I therefore pick up on some of the unresolved aspects mentioned in the previous section (Section 5.2), and discuss several possible enhancements of missing data approaches (5.3.1). Subsequently, the results from the present dissertation are evaluated in terms of generalizability to other testing situations (5.3.2). Possible extensions to further research areas are illustrated.

5.3.1 Advances in accounting for missing data

Model-based approaches aim at modeling the missing data process. The reason for a missing value gives valuable information on this missing data process. So far, not all aspects influencing the missing propensity are identified. The threshold for omitting an item might be higher for some people than for others. Possible relevant factors include context related variables such as motivation (see, e.g., Wise & DeMars, 2005), personality traits such as risk-averseness, and content related variables such as self-efficacy and self-concept in the assessed competence area. These features are seldom measured as out-of-test variables, and could be investigated in future studies. To further verify that the missing propensity is a rather person-specific characteristic, studies that examine the missing propensity across several time points could be conducted. Some large-scale studies have longitudinal designs, which provide a great opportunity to investigate changes in the omission behavior and changes in the amounts of not-reached items over time.

Process data from computerized tests might provide additional valuable information on why an examinee failed to give an answer to a particular item. Currently, researchers explore action sequences and response time data in order to gain information on why an item has been omitted (He & von Davier, 2015; Weeks, 2015). Action sequences refer to sequences of

examinee behavior within a test, which can be obtained from computer log files. Both approaches enhance the accuracy of determining whether the examinee could have answered the item correctly. For example, if the examinee barely took any time to work on the item before skipping it, chances are high that he or she would have answered it correctly upon thorough consideration. Treating such speedy omissions as incorrect responses is hardly justifiable (Weeks, 2015). Action sequences shed light on examinee engagement with a respective item. This method is especially useful with regard to more interactive items types, which become more common in order to, for example, assess technological and computer skills (see, e.g., the Programme for International Assessment of Adult Competencies [PIAAC]; Schleicher, 2008). The items involve clicking several buttons or links before the examinee can provide the correct answer. If the test taker hardly interacts with the item, chances are high he or she was rather unmotivated to solve the item correctly (He & von Davier, 2015). This type of skipping behavior should thus rather be considered an omission (i.e., a not-administered item) than an incorrect item (He & von Davier, 2015).

Context analyses and process data provide indications for relevant explaining variables and for why an item might have been skipped. However, the researcher still remains oblivious as to the true reason for a nonresponse on a particular item by a particular examinee. Existing research could be enhanced by a qualitative study—supplementary to the actual competence assessment—in which examinees elaborate on why they did not provide an answer for the items they skipped or why they failed to reach the end of the test (Glynn et al., 1993). It could also include giving examinees additional time to work on the items they had not responded to, asking them what their response would have been (Jakwerth et al., 1999). Such a study would offer valuable information on the missing data mechanism. Jakwerth et al. (1999) conducted this type of study in the NAEP assessment. However, they refrained from combining the qualitative

information with quantitative methods. Since qualitative studies are rather extensive and time consuming, they cannot be conducted complementary to each data assessment. It would thus be beneficiary if cues in the quantitative data allowed inferences on the true reason for an omission. Qualitative information could, for example, be combined with process data approaches and aid in determining which type of process data adequately indicates the reason for a missing value. The qualitative study may thus validate and enhance previous attempts on modeling the missing data mechanism. At large, a model that more efficiently describes the missing process might be developed, and missing data could thus be more accurately handled in empirical analyses.

Another possible line of research already addressed in the previous section concerns examinee behavior under different scaling conditions. A study in which all test takers receive the same items but different instructions might be very informative. The instructions could vary in what people are told about the scaling method, and might include a specific description of how missing values are handled as well as advice on the respective optimal test-taking strategy. This could shed light on test-taking behavior with regard to different missing data approaches, its influence on the occurrence of omitted and not-reached items, and its impact on the scaling of competencies.

All information from different studies on the missing data mechanism should be included in order to precisely determine under which conditions which missing data approaches result in accurate parameter estimates. The methods presented in the current dissertation valuably enhanced previous approaches, and our findings provide pertinent points of reference for future investigations. Simulation studies might incorporate our results regarding (a) the skewly distributed omission propensity, (b) multidimensionality of the omission process with regard to different item response formats, and (c) explaining variables. These considerations allow for realistic data sets in which the missing data mechanism more accurately maps the missing data

mechanism in real competence tests. Other plausible methods for inducing missing values are thinkable and should be considered, for example, letting the probability for a missing value depend on the true item response (see, e.g., Culbertson, 2011; de Ayala et al., 2001; Rose et al, 2010). Additionally, investigations that simultaneously consider the propensity to omit and the propensity to not reach items make for possible interesting extensions (Pohl et al., 2014; Rose, 2013).

5.3.2 Generalizability of results

The studies of the present dissertation were conducted using competence test data from a national German low-stakes assessment. It would be interesting to replicate our findings on the missing data mechanism with data from other national or international low-stakes assessments. Furthermore, our approaches could be extended to the scaling of noncognitive skills such as self-concept or perseverance. These traits are also considered relevant with regard to educational growth, and are frequently assessed alongside cognitive skills. It would be interesting to investigate the properties and features of the missing data mechanism in these areas. Tests on noncognitive skills might profit from the model-based approach, especially if the missing responses depend on the underlying trait. With regard to high-stakes assessments in education, the process underlying missing data probably differs from that in low-stakes assessments. Examinees most likely show a greater degree of ambition, since their individual test score impacts their future educational or academic career. Nonresponse might predominantly be due to not knowing the correct answer, thus incorrect scoring would possibly be the appropriate missing data method for high-stakes assessments. These considerations require future investigation. Certainly, model-based approaches could also be applied to these types of tests, though the missing data mechanism would need thorough examination prior to their application.

Besides testing in educational large-scale assessment, data analysts in other areas might profit from applying model-based approaches, especially if the nonresponse highly depends on the underlying trait. Missing data present potential problems to the validity of any empirical study. If missing values occur systematically, the inferences drawn from only the observed data might deviate from the inferences one would have obtained if the data had been complete (Rubin, 1976). Research conducted in personality assessment, for example, may benefit from model-based approaches. The missing data pattern in personality questionnaires probably holds valuable information on the trait that is being assessed, for example in cases where people with higher expressions of the personality trait also tend to skip more items. This information could be included in the measurement model for the personality trait, thus accounting for questions the examinee refused to respond to. Keep in mind that including the missing data mechanism requires thorough investigation of the missing data process, since it needs to be accurately modeled. For this, the methods presented in the current dissertation could be applied and enhanced.

5.4 Conclusion

The present dissertation valuably contributes to missing data research in the context of large-scale assessment studies. Missing values can influence parameter estimation on various levels, and an inadequate treatment might result in incorrect inferences from data analysis. We showed that the hitherto used methods are inferior to model-based approaches, and that their implementation in the scaling of competence tests improves accuracy of parameter estimates. Results from empirical analyses demonstrated the applicability of model-based approaches to competence test data of large-scale assessments. We also illustrated possible modifications and enhancements in order to more accurately take the missing data mechanism into account. Our studies on features of the missing process give a better understanding of how and why missing

values occur, thus allowing a deeper insight into the missing data mechanism. This information might aid in optimizing missing data approaches. Several questions regarding a state-of-the-art missing data method for large-scale assessment studies remain unanswered, and need to be considered in future investigations. A change in the scaling method of several large-scale assessments lies certainly ahead. The challenge to present superior alternatives convincingly enough remains. It would be desirable to implement a method that ensures that missing values can, in fact, be considered a nuisance, and not one of the main reasons for unreliable test results.

6 References

Adams, R., & Wu, M. (2002). *PISA 2000 Technical Report*. Paris: OECD.

Ainley, J., Fraillon, J. & Freeman, C. (2008). *National Assessment Program. ICT Literacy Years 6 and 10 Report 2005.* Australian Council for Educational Research (ACER), Camberwell Victoria: ACER.

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T. & Daugherty, R. (2011). *Policy Effects of PISA.* Oxford University Centre for Educational Assessment (OUCEA) Report. Retrieved from http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology in Education, 16,* 441-462.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983–84 Technical Report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I., Rock, D., & Yamamoto, K. (1992). An experiment in monetary incentives. *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 393–398. Alexandria, VA: American Statistical Association.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35,* 321–364. doi: 10.1207/S15327906MBR3503_03.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental rest scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5–17. doi: 10.1007/s11618-011-0178-3

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*, 21-33.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Auflage). Berlin: Springer.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly, 57*, 62–79.

Culbertson, M. (2011). *Is it wrong? Handling missing responses in IRT.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, USA.

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*, 213-234. doi: 10.1111/j.1745-3984.2001.tb01124.x

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, 39*, 1–39.

Department of Education, Early Learning and Culture (2012). *Preparing students for PISA - Mathematical Literacy - Teacher's handbook.* Prince Edward Island: EECD. Retrieved from http://www.gov.pe.ca/photos/original/ed_PISA_math1.pdf

DePascale, C.A. (2003). The ideal role of large-scale testing in a comprehensive assessment system. *Journal of Applied Testing Technology, 5*, 1-11. Retrieved from http://www.testpublishers.org/assets/documents/volume%205%20issue%201%20The%20ideal%20role.pdf

Embretson, S., & Reise, S. (2000). *Item Response Theory for psychologists.* Mahwah, NJ: Erlbaum.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225-245. doi:10.1111/j.1745-3984.2008.00062.x

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. Chapter 11 in J.F. Olson, M.O. Martin, and I.V.S. Mullis (Eds.), *TIMSS 2007 Technical Report* (pp. 225-280). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907-922. doi: 10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (in press). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling.*

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples*. New York, NJ: Springer. doi: 10.1007/978-1-4612-4976-4_10

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with followups. *Journal of American Statistical Association, 88,* 984–993.

Haberman, S. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (ETS Research Report RR-13-32). Princeton, NJ: Educational Testing Service.

He, Q., & von Davier, M. (2015). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara & M. Mosharraf (Eds.), *Handbook of Research on Computational Tools for Real-World Skill Development*, IGI Global.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, limited dependent variables, and simple estimator for such models. *Annals of Economic and Social Measurement, 5,* 475-492.

Hohensinn, C., & Kubinger, K. D. (2011). Applying Item Response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71,* 732- 746. doi: 10.1177/0013164410390032

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,* 1-17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., Stancavage, F. B., & Reed, E. D. (1999). *An investigation of why students do not respond to questions*, NAEP Validity Studies, Working Paper Series, American Institutes for Research, Palo Alto, CA.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 Technical Report* (Rep. No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Katsikatsou, M., Kuha, J., & Moustaki, I. (2014). *Modelling missing values in cross-national surveys: a latent variable approach.* Workshop on Cross-National Surveys: Methods of Design and Analysis.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Tech. Rep. No. 357). Los Angeles: Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264. doi: 10.1007/BF02291471

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 Technical Report.* Chestnut Hill, MA: Boston College.

Martin, S., Helmschrott, S., & Rammstedt, B. (2014). The use of respondent incentives in PIAAC: The field test experiment in Germany. *methods, data, analyses, 8*, 223-242. doi: 10.12758/mda.2014.009

Matters, G., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement, 63*, 239-256.

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Molenaar, D. (2007). Accounting for non-normality in latent regression models using a cumulative normal selection function. *Measurement and Research Department Reports, 3*. Arnhem: Cito.

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*, 259-276.

OECD (2012). *PISA 2009 Technical Report, PISA*, OECD Publishing.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A, 162*, 177-194. doi: 10.1111/1467-985X.00129

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208.

PISA Project Consortium (2003). *PISA 2003 main study test administrator's manual.* Paris: OECD. Retrieved from http://www.acer.edu.au/files/pisa2003_test_adminitrator_manual.pdf

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report - Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in Item Response Theory models. *Educational and Psychological Measurement, 74*, 423–452. doi: 10.1177/0013164413504926

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Nielsen & Lydiche. (Expanded edition, 1980)

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89,* 846–866.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement.* Ph.D. thesis, Friedrich-Schiller-University Jena, Dept. of Methodology and Evaluation Research.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report ETS RR-10-11), Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sabers, D. L., & Feldt, L. S. (1968). An empirical study of the effect of the correction for chance success on the on the reliability and validity of an aptitude test. *Journal of Educational Measurement , 5*, 251-258.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147−177.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association, 95,* 144–154. doi: 10.1080/01621459.2000.10473910

Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education, 54*, 627-650.

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505–528. doi: 10.1207/s15327906mbr3804_4

Stephen, M. (2013, December 02). PISA: Poor academic standards – and an even poorer test. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/

Stocking, M. L., Eignor, D., & Cook, L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true score equation procedures* (RR-88-41). Princeton, NJ: Educational Testing Service.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of multilog. *Applied Psychological Measurement, 16*, 1-6. doi: 10.1177/014662169201600101

von Davier, M. (2001). WINMIRA 2001 [Computer software].

von Davier, M. (2005a). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (2013). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. Dordrecht: Springer.

Weeks, J. P. (2015). *Using response time data to inform the coding of omitted responses*. Presentation at the Educational Testing Service (ETS), Princeton, NJ.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17. doi: 10.1207/s15326977ea1001_1

Zhang, J. (2013). *Relationships between missing response and skill mastery profiles of cognitive diagnostic assessment.* Ph.D. thesis, University of Toronto, Dep. of Curriculum, Teaching, and Learning.

Zwinderman, A. H. & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the rasch model. *Applied Psychological Measurement, 14*, 73-81. doi:10.1177/014662169001400107

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Dissertation selbstständig und ohne unerlaubte Hilfe angefertigt zu haben. Bei der Verfassung der Dissertation wurden keine anderen als die im Text aufgeführten Hilfsmittel verwendet. Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen Hochschule oder bei einem anderen Fachbereich beantragt.

Bamberg, 30.09.2015