*Article*

Educational and Psychological Measurement
2014, Vol. 74(3) 423–452

DOI: 10.1177/0013164413504926
epm.sagepub.com
SAGE

# Dealing With Omitted and Not-Reached Items in Competence Tests: Evaluating Approaches Accounting for Missing Responses in Item Response Theory Models

**Steffi Pohl[1], Linda Gräfe[2], and Norman Rose[3]**

## Abstract

Data from competence tests usually show a number of missing responses on test items due to both omitted and not-reached items. Different approaches for dealing with missing responses exist, and there are no clear guidelines on which of those to use. While classical approaches rely on an ignorable missing data mechanism, the most recently developed model-based approaches account for nonignorable missing responses. Model-based approaches include the missing propensity in the measurement model. Although these models are very promising, the assumptions made in these models have not yet been tested for plausibility in empirical data. Furthermore, studies investigating the performance of different approaches have only focused on one kind of missing response at once. In this study, we investigated the performance of classical and model-based approaches in empirical data, accounting for different kinds of missing responses simultaneously. We confirmed the existence of a unidimensional tendency to omit items. Indicating nonignorability of the missing mechanism, missing tendency due to both omitted and not-reached items correlated with ability. However, results on parameter estimation showed that ignoring missing

[1]University of Bamberg, Bamberg, Germany
[2]University of Jena, Jena, Germany
[3]University of Tübingen, Tübingen, Germany

**Corresponding Author:**
Steffi Pohl, National Educational Panel Study, University of Bamberg, Wilhelmsplatz 3, Bamberg 96045, Germany.
Email: steffi.pohl@uni-bamberg.de

responses was sufficient to account for missing responses, and that the missing propensity was not needed in the model. The results from the empirical study were corroborated in a complete case simulation.

**Keywords**

missing responses, item response theory, nonignorability, missing propensity

## Introduction

### *Missing Responses in Competence Tests*

Large-scale assessments, such as the Program for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the National Educational Panel Study (NEPS) aim at accurately measuring competencies such as reading comprehension or mathematical literacy. Competence test data usually include missing responses on test items. Missing responses may be due to (a) items that are not administered, (b) omitted items, or (c) items that are not reached because of test time limitations. The amount of missing responses in large-scale assessments is not negligible. In PISA 2006, for example, across all countries and all three domains (mathematics, reading, and science), an average of 10% (in Germany 8.37%) of the items were omitted and 4% (in Germany 1.15%) were not reached (Organisation for Economic Co-operation and Development [OECD], 2009, pp. 219-220). For mathematics and science in TIMSS 2003, an average of 3.73% of the items were not reached in Grade 8 and 5.96% in Grade 4 (Mullis, Martin, & Diaconu, 2004, p. 249). In the NAEP mathematics assessment of 1986, the researchers had to deal with a very high rate of not-reached items: 41% of the test items had at least 20% missing responses due to test time limits (Johnson & Zwick, 1990). Because of an extended testing time in 1990, the amount decreased to 8% of items with 20% missing responses due to items that were not reached (Koretz, Lewis, Skewes-Cox, & Burstein, 1993). In the same test, 5%-9% of the items (depending on the grade) had omission rates above 10% (Koretz et al., 1993). While not-administered items are usually missing completely at random (MCAR), omitted and not-reached items are usually nonignorable and may lead to biased estimates of item and person parameters (see, e.g., Lord, 1974; Mislevy & Wu, 1996). In order to avoid biased item and person parameter estimates, the missing responses need to be appropriately accounted for.

### *Dealing With Missing Responses*

There are different approaches of dealing with missing responses. They may be classified in classical approaches, imputation-based approaches, and model-based approaches for nonignorable missing data.

*Classical approaches.* In the classical approaches (see, e.g., De Ayyala, Plake, & Impara, 2001 or Rose, von Davier, & Xu, 2010), missing values are accounted for by either scoring or ignoring the missing responses in the estimation. There are different ways of dealing with missing responses: (a) Missing responses may be ignored and, thus, treated as if they were not administered. This approach assumes that missing responses are missing at random (MAR), given the observed responses on the items in the test and other covariates in the background model. (b) Missing responses may also be scored as incorrect responses, assuming that the subject does not know the answer. This is a deterministic scoring approach ignoring the fact that the subject has a positive probability to solve the item, given its individual trait level. Lord (1974) showed that the incorrect scoring method results in biased parameter estimates and proposed (c) to score missing responses as fractional correct, for example, by scoring them according to the probability of guessing correctly (e.g., a score of 0.25 for a multiple choice item with four response options). Further alternative fractional correct scoring methods have been proposed and their performance has been investigated by Culbertson (2011), for example. (d) In some educational studies (e.g., PISA, TIMSS, and PIRLS), a two-stage procedure for treating missing responses is used (see, e.g., OECD, 2009). For the estimation of the item parameters, missing responses are ignored. The estimated item parameters are then used as fixed parameters for the estimation of person parameters where missing responses are scored as incorrect.

Classical approaches are the most common approaches for treating missing responses in large-scale educational assessments. In PISA, TIMSS, and PIRLS, missing responses due to both omitted and not-reached items are ignored for item parameter estimation and scored as incorrect for person parameter estimation (see, e.g., OECD, 2009; Macaskill, Adams, & Wu, 1998; Martin, Mullis, & Kennedy, 2007). In NAEP (Allen, Donoghue, & Schoeps, 2001), different approaches are used for different kinds of missing responses. In item and person parameter estimation, those due to omitted items are scored as fractional correct while those due to not-reached items are ignored.

De Ayyala et al. (2001), Finch (2008), Lord (1974), Ludlow and O'leary (1999), as well as Rose et al. (2010) investigated the performance of different strategies to account for missing responses. They showed that scoring missing responses as incorrect (Methods B and D) results in biased parameter estimates and overestimated reliability while ignoring missing responses (Method A) results in unbiased parameter and reliability estimates. Scoring missing responses as fractional correct (Method C) is almost as accurate as using complete data. Ignoring missing responses has many advantages. It is easy to implement and so far, studies confirm that it results in unbiased parameter estimates. However, ignoring missing responses in the estimation draws on the assumption of ignorability. If ignorability does not hold, the approach may not be appropriate.

*Imputation-based approaches.* There are different ways of imputing missing responses on test items. These include corrected mean substitution (Bernaards & Sijtsma, 2000;

Huisman & Molenaar, 2001; Sijtsma & van der Ark, 2003), response function imputation (Sijtsma & van der Ark, 2003), expectation-maximization algorithm (Little & Rubin, 2002), and multiple imputation (MI; Rubin, 1987). Finch (2008) investigated the performance of the different imputation-based methods and found that MI outperforms all other imputation methods. MI uses the valid responses, the responses of other participants with similar characteristics, and observed information on covariates, when a background model is included, to impute the missing responses on the test items. MI is a Bayesian approach in which the imputed values are drawn from a posterior distribution. The imputation is done multiple times in order to get a correct estimate of the standard error of the parameter.

Using MI to deal with missing responses in item response theory (IRT) models has a substantial disadvantage. For statistical inference based on person parameter estimates, plausible values have proved to be most appropriate (Mislevy, 1991, 1993). However, plausible values themselves are multiple imputations of latent variables. Hence, generating plausible values from multiple imputed data sets requires nested MI (Harel & Schafer, 2003; Reiter & Raghunathan, 2007; Shen, 2000). A large number of plausible values result from nested MI whose generation requires repeated parameter estimation of computationally demanding IRT models. Therefore, nested MI becomes impractical in measurement models used for large-scale assessments. For these reasons, multiple imputation is not frequently used for dealing with item nonresponse in competence tests.

*Model-based approaches for nonignorable missing data mechanisms.* Recently, model-based approaches for dealing with nonignorable missing data in IRT models were developed. These approaches account for the fact that the missing data mechanism might be nonignorable. In fact, Mislevy and Wu (1988, 1996) provide arguments for the conclusion that the missing mechanism for omitted and not-reached items is nonignorable. Ignorability for Bayesian inferences is given when MCAR or MAR and distinctness hold (Rubin, 1976). Missing responses due to test time limits are—if items are not ordered by difficulty—independent of the actual response of the subject. This is different for omitted items. As studies (e.g., Stocking, Eignor, & Cook, 1988) show, omission more likely occurs when the participant does not know the answer to the item and, thus, MAR is a less plausible assumption. Furthermore, different studies (Glas & Pimentel, 2008; Holman & Glas, 2005; Matters & Burnett, 2003; Rose et al., 2010; Sheriffs & Boomer, 1954; Stocking et al., 1988) demonstrated that ability and missing propensity are not distinct. Instead, missing responses due to omission and test time limits often depend on the ability of the person. The more able a person, the smaller the number of missing responses in tests (latent correlations up to $-0.45$).

As a consequence of possible nonignorability, we must draw inferences from the full model for the joint distribution of the missingness and the item responses (Lord, 1983). In the model-based approaches, the missing tendency is included in the IRT model and accounted for in the estimation of the item and person parameters. The

missing tendency is included either (a) via a latent missing propensity (Holman & Glas, 2005; Glas & Pimentel, 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999) that is accounted for in a multidimensional IRT model (*latent approach*) or (b) by modeling a manifest missing variable that is accounted for in a latent regression or multiple-group IRT model (*manifest approach*, see, e.g., Rose et al., 2010). While the manifest approach is applicable to both omitted and not-reached items, in the latent approach there are different models for both types of missing responses. Holman and Glas (2005) proposed (a) a model accounting for omitted items while Glas and Pimentel (2008) developed (b) a model accounting for items that were not reached. In all model-based approaches, there is a unidimensional IRT measurement model for the responses $Y_i$ of the participants on item $i$. Missing responses due to omitted and not-reached items on the observed responses $Y_i$ of item $i$ are treated as missing values in the measurement model of $Y_i$.

*1a. The latent approach for modeling missing responses due to omitted items.* Based on work by Moustaki and Knott (2000) and O'Muircheartaigh and Moustaki (1999), Holman and Glas (2005) proposed three models accounting for omitted responses. These three models are statistically equivalent; however, they do differ in the meaning of the latent variables (Rose et al., 2010). The model with the most straightforward interpretation of the latent variables is depicted in Figure 1a. Observed responses on the test items $i$ are denoted by $Y_i$ and the latent ability to be estimated is denoted by $\xi$. The manifest missing indicators $d_i$ in the model are computed as

$$d_i = \begin{cases} 0 & \text{if } Y_i \text{ is observed} \\ 1 & \text{if } Y_i \text{ is omitted.} \end{cases} \quad (1)$$

A latent variable $\theta$, which may be interpreted as the propensity of a person to omit an item, is then modeled based on these manifest missing indicators. Thus, there is a separate measurement model for missing responses and observed scores, and a latent competence $\xi$ as well as a latent missing propensity $\theta$ are modeled.

In their article, Holman and Glas (2005) introduced the model using a 2PL (two-parameter logistic) model for both latent variables. They noted, however, that a 1PL model may be more convenient for the measurement model of the missing indicators. Of course, one can test whether a 1PL or a 2PL model is needed for a specific test. An implicit assumption of the model is that the missing indicators fit a unidimensional measurement model. Since the items of a test are constructed in a way that the *responses* $Y_i$ meet a unidimensional IRT model but not that the *missing indicators* $d_i$ meet a unidimensional IRT model, this assumption is not trivial.

As Rose (2013) demonstrated, the latent approach potentially fails to correct for nonignorable item nonresponses if a possible multidimensionality of the latent response propensity is not taken into account. It may well be that the missing indicators are best described by a multi-dimensional model, for example, if some participants are more likely to omit mathematics items dealing with geometry and others omit items dealing with algebra. The missing indicators may not, then, correlate very high at all, and a unidimensional measurement model may not hold. Whether the

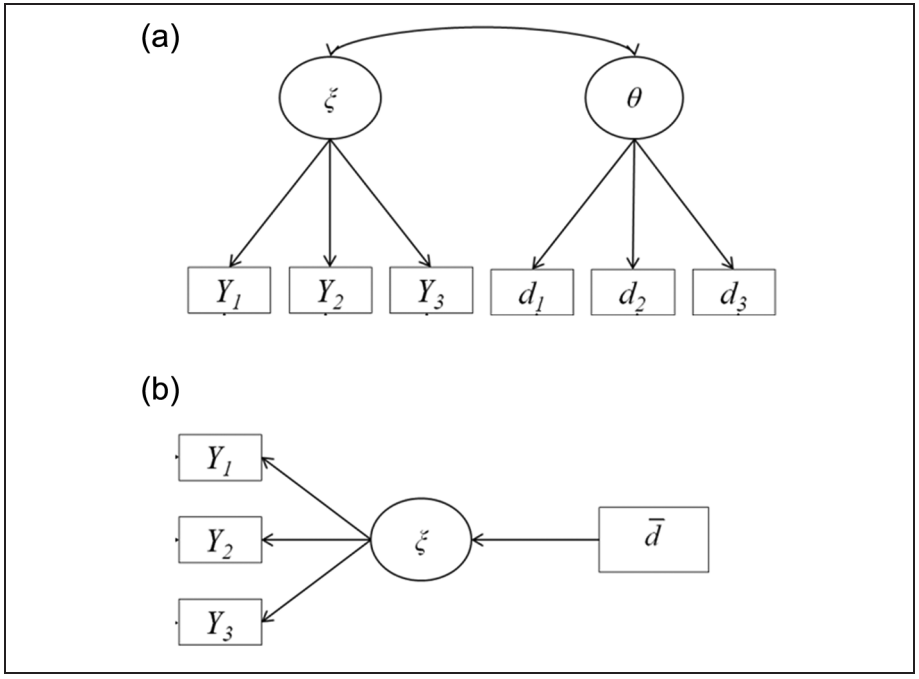


**Figure 1.** Model-based approaches: Including the missing propensity as (a) a latent or (b) a manifest variable.

measurement model for the missing indicators holds in practice needs to be investigated.

*1b. The latent approach for modeling missing responses due to not-reached items.* Glas and Pimentel (2008) adjusted the model by Holman and Glas (2005) to also account for missing responses due to not-reached items. As for omitted items, a latent variable is specified for the missing indicators. The latent missing propensity for not-reached items represents the speed component. The adjustments account for the fact that missing responses due to not-reached items have some specific features—for example, that the first item is always reached and that the number of missing responses due to not-reached items increases with the position of the item in the test. The manifest missing indicators $d_i$ for not-reached items are computed as

$$d_i = \begin{cases} 0 & \text{if } Y_i \text{ is observed} \\ 1 & \text{if } Y_i \text{ is the first item that is not reached} \\ \text{NA} & \text{otherwise} \end{cases} \quad (2)$$

with NA indicating a missing value. Linear restrictions based on the position of the item in the test are imposed on the difficulties $\beta_{di}$ of the manifest variables $d_i$:

$$\beta_{di} = \tau_0 + (k - K)\tau_1, \tag{3}$$

where $K$ is the number of items in the test and $k$ is the position of the item in the test. $\tau_0$ and $\tau_1$ are to be estimated and indicate, respectively, the difficulty of the missing indicator of the last item of a test and the linear decrease in difficulty for each position earlier in the test.

The linearity assumption does not necessarily need to hold in empirical studies. In reading tests, when items are clustered within texts, the difference in difficulty between the last item of a text and the first item of the following text may be larger than the difference between two adjacent items of the same text. A function that also estimates such steps after each text may better represent the data. This, however, needs to be tested in empirical data. Note that since not-reached items result in a monotone missing pattern, the measurement model may only fail to fit the data because of false restrictions on the item difficulties.

*2. The manifest approach for modeling missing responses.* The latent approaches for modeling missing responses may result in estimation problems (see, e.g., Cai, 2010), especially when the sample size and the number of missing responses are rather small or when both the latent ability and the latent response propensity are multidimensional (Rose, 2013; Rose & von Davier, 2013). Modeling a manifest instead of a latent missing propensity may solve these estimation problems. In the manifest approach, the relative number of missing responses (for either kind of missing response) in a test is computed for each person as

$$\bar{d} = \frac{1}{K}\sum_{i=1}^{K} d_i \tag{4}$$

with $i$ indicating the item and $K$ the total number of items in the test. $d_i$ denotes a missing indicator as defined in Equation 1 (for not-reached items $d_i = 1$ when $Y_i$ is not reached). Instead of a latent missing propensity in the manifest approach, a manifest variable representing the relative number of missing responses is included in the measurement model (see Figure 1b). The manifest missing variable is incorporated in the model via latent regression (or alternatively, a multiple-group IRT model; see Rose et al., 2010). The relative number of missing responses in the manifest approach is accounted for in the estimation of the item parameters and ability scores. There is no restriction on the kind of measurement model for the latent ability.

The manifest and the latent approach for modeling missing responses differ in some aspects. In the latent approach, the existence of a latent missing propensity is assumed and the model itself is more complex, needs more assumptions, and may lead to convergence problems when the number of missing responses is small. The manifest approach is much easier to implement and to estimate. In the manifest approach it is—averaging across the missing indicators—implicitly assumed that there is a unidimensional missing propensity. This assumption is, however, not tested in the model. Furthermore, including a fallible measure of the missing propensity

instead of a latent one may distort the correlation and, thus, a result in a less-efficient bias reduction (e.g., Campbell & Erlebacher, 1975; Lord, 1960).

## *Performance of Model-Based Approaches*

Holman and Glas (2005) as well as Glas and Pimentel (2008) showed that model-based approaches for nonignorable missing data considerably reduce bias caused by ignoring the missing-data mechanism. Rose et al. (2010) compared the performance of model-based approaches with classical approaches for omitted items in a simulation study and validated their results in an empirical analysis using PISA data. They showed that including the missing propensity does result in unbiased estimates of item and person parameters. In terms of bias of parameter estimates, in their simulated data example, the model-based approaches performed as equally well as the approach of ignoring missing responses in the estimation. All these methods correctly estimate the reliability. Model-based approaches do, however, result in higher reliability estimates.

The approach of ignoring missing responses relies on the assumption that the missing responses are ignorable. The model-based approaches on the contrary deal with nonignorable missing data mechanisms. The correlation between ability and missing propensity, which can be estimated in the model-based approaches, indicates the amount of nonignorability (Glas & Pimentel, 2008). Checking whether parameter estimates issued from an approach of ignoring missing responses and a model-based approach are the same can be viewed as a test of the robustness of the approach ignoring missing values to violations of the ignorability assumption. If the model-based approaches do result in different parameter estimates than ignoring missing responses, the assumption of ignorability may be falsified. Since model-based approaches have only recently been developed, not much is known about the need for including the missing propensity in applications.

Nonignorability is given when ability correlates with the missing tendency. In this case, it may be necessary to jointly model the responses and the missing indicators. The relationship estimated between missing propensity and ability may be differently affected by attenuation bias when modeling a manifest instead of a latent missing propensity, and this in turn may affect parameter estimation. So the question is not only whether the missing propensity is needed in the model but also whether it needs to be included as a latent variable or whether it is sufficient to account for it in a manifest form.

Last, but not least, the performance of the model-based approaches will depend on the plausibility of the model assumptions. The unidimensionality of the omission indicators is explicitly assumed in the latent approach and implicitly assumed in the manifest approach. Averaging the number of missing responses across items is only justified if a unidimensional construct is supported. Rose (2013) noted that model-based approaches may result in biased parameter estimates when the unidimensionality of the missing propensity does not hold. Up to now, the plausibility of the

unidimensionality assumption of the missing indicators has not been tested in empirical data.

### *Accounting for Different Types of Missing Responses*

Studies investigating the performance of the different approaches for dealing with missing responses have either focused only on one kind of missing response (e.g., De Ayala et al., 2001; Culbertson, 2011; Finch, 2008; Glas & Pimentel, 2008; Holman & Glas, 2005) or not distinguished between different kinds of missing responses (e.g., Rose et al., 2010). However, in data from competence tests, usually both kinds of missing responses are prevalent and the mechanism underlying the missing process will most likely be different for omitted versus not-reached items. Previous studies (e.g., Stocking et al., 1988; van den Wollenberg, 1979) suggest that the tendency to omit items depends on the difficulty of the item and the ability of the test taker. The amount of not-reached items may, on the other hand, depend on the test-taking strategy, the motivation, or the ability of the test taker (e.g., Mislevy & Wu, 1988). Since different processes are involved, the different missing responses may need to be modeled separately, accounting for each kind of missing response in a different way. So far, although different types of missing responses occur in competence tests, not much research has investigated the performance of the approaches for both kinds of missing responses at once.

## Research Questions

In this article, we investigate which approach best accounts for missing responses in competence tests. Contrary to previous work, we consider both kinds of missing responses—missing responses due to omitted and due to not-reached items—in one model and try to find appropriate models accounting for both of them. We especially focus on the model-based approaches since they may account for nonignorable missing responses.

As described above, the performance of the model-based approaches will also rely on the appropriateness of their assumptions. Therefore, we first investigate the appropriateness of the unidimensionality assumption of the omission indicators. Research questions regarding this model assumption are as follows: Is there a unidimensional missing propensity for the whole test? Or is the omission tendency item-specific?

Whether the missing responses are ignorable (or the respective models are robust against violations of ignorability) or whether the model-based approaches are needed to account for nonignorability also depends on the relationship between missing propensity and ability. If missing propensity and ability are related, missing indicators and item responses may need to be modeled jointly. Another research question, therefore, focuses on the relationship of the missing propensity with ability: Is the missing propensity related to ability? How much does this relationship differ when modeling a manifest instead of a latent missing propensity?

Finally, we take interest in the performance of the different approaches regarding item and person parameter estimation. Research questions considering performance are: Are the missing responses due to omitted and not-reached items robust against violations of ignorability or do we need to use models for nonignorable missing responses? Is there a difference in including a latent or a manifest missing propensity? How do the model-based approaches perform compared with commonly implemented methods?

## Method

### *Data*

The National Educational Panel Study (NEPS) is a large-scale study with a longitudinal design that aims at investigating the development of competencies of German inhabitants across the whole life span—from infancy to old age. In the NEPS, researchers draw a representative sample that is tested according to a multicohort sequence design (Blossfeld, Roßbach, & von Maurice, 2011; Blossfeld, von Maurice, & Schneider, 2011). In the NEPS, there are six starting cohorts: infants, kindergartners, fifth graders, ninth graders, first-year college students, and adults. Each cohort consists of a sample size between $N = 3{,}000$ and $N = 15{,}000$ people. In each cohort, different competencies are assessed repeatedly across the life span. These can be grouped in domain-general cognitive functions, domain-specific cognitive competencies, as well as meta and social competencies (Weinert et al., 2011). Domain-specific cognitive competencies include reading, mathematics, and scientific competence and are measured coherently across the whole life span. Testing one competence at one measurement occasion lasts about 30 minutes and includes 20 to 40 items. The items are developed to fit a Rasch (Rasch, 1960) or a partial credit model (Masters, 1982). A detailed description of the scaling of the competence tests in the NEPS can be found in Pohl and Carstensen (2012a).

For the analyses in this article, we used data from the reading and mathematical competence tests of fifth-grade students. Reading and mathematics are basic competence domains and are the first competence domains tested in most of the NEPS cohorts. Furthermore, the two competence domains showed some differences in the occurrence of missing responses. While the reading test showed the largest amount of not-reached items (Pohl, Haberkorn, Hardt, & Wiegand, 2012), the mathematics test showed only a small amount of not-reached items (Duchhardt & Gerdes, 2012). Using these competence domains allowed us to test the models on a spectrum of real data.

Both tests contained a number of tasks with different item formats. Most of the items were simple multiple-choice (MC) items with one correct response out of four response options. Other item formats included complex MC items, matching items, and short-constructed responses. Complex MC items consist of a common stimulus with several subtasks that need to be answered by either agreeing or disagreeing with a given statement. Matching items require the test taker to sort a number of responses to a given set of statements; they are usually applied to sort titles to paragraphs of a text and are only used for reading tests. Open constructed responses are used for

mathematics tests. They require the test taker to give a short answer—for example, a number. In estimating item and person parameters in the NEPS, complex MC items and matching items are treated as polytomous items indicating the number of correctly answered subtasks (see Pohl & Carstensen, 2012a).

In the present analyses, each subtask was treated as a single dichotomous item. This allowed us to more thoroughly investigate the missing responses, including those that occurred within complex MC and matching items.[1] In total, there were 28 items for mathematics and 59 items (including subtasks) for reading. Because of perfect local dependence of three matching items, three subtasks of matching items had to be excluded from the analyses (see Pohl & Carstensen, 2012a), resulting in 56 items. While the items in the mathematics test did not share a common stimulus, the reading test consisted of five texts, each with a set of items referring to the same text.

A total of 4,989 subjects took the reading test. A total of 2,504 took the reading test before the mathematics test and 2,485 took the reading test after the mathematics test. Thirteen people were excluded from the analyses because they had less than three valid responses. The remaining 4,976 subjects were included in the analyses. From the 5,208 subjects who took the mathematics test,[2] 14 subjects were excluded from the analyses because of a limited number of valid responses. In all, $N = 5{,}194$ entered the analyses.

Looking at the degree of missing responses, the average rates of omitted items per person in reading and mathematics were very similar (5.37% and 5.15%, respectively). However, focusing on the not-reached items, there were on average more missing responses in the reading test (13.46%) than in the mathematics test (1.32%). That led to differences in the total amount of missing responses between the domains, which were on average 18.82% for reading and 6.47% for mathematics (see Pohl et al., 2012 as well as Duchhardt & Gerdes, 2012 for a more detailed description of the missing responses in the reading and mathematics tests, respectively).

## Analyses

In the following analyses, we distinguish between different kinds of missing responses. In order to disentangle the different missing mechanisms corresponding to the different kinds of missing responses, we first evaluated the models dealing with (a) omitted items only or (b) not-reached items only. Then we analyzed both kinds of missing responses simultaneously, (c) using a composite across both types of missing responses or (d) dealing with both kinds of missing responses separately in one model. Note that when only one kind of missing response was regarded, the other kind of missing response (which still occurred in the data) was ignored in the estimation. For (c), a composite of both kinds of missing responses was constructed based on missing indicators for each item $i$ being defined as

$$d_i = \begin{cases} 0 & \text{if } Y_i \text{ is observed} \\ 1 & \text{if } Y_i \text{ is omitted or not reached.} \end{cases} \quad (5)$$

The manifest missing propensity based on the composite missing indicators was then defined as the mean of the respective $d_i$. The latent missing propensity was modeled based on the missing indicators defined in Equation 5. The manifest and the latent missing propensity of the composite score represented the general tendency for missing responses that comprised omission and speed. As the missing mechanism may well have been different between the two kinds of missing responses, in the last procedure they were separately accounted for in the same model. This was done by including both missing propensities in the model. The respective model, including manifest missing propensities is depicted in Figure 2a, and the model, including latent missing propensities is depicted in Figure 2b. The missing indicators for the omitted items in these models were defined as in Equation 1 while the missing indicators for not-reached items were defined as in Equation 2. While the manifest missing propensities for both types of missing responses were constructed as the average across the respective missing indicators (Equation 4), the latent missing propensities were specified as proposed by Holman and Glas (2005) for omission and by Glas and Pimentel (2008) for speed. All analyses were performed using ConQuest (Wu, Adams, & Haldane, 2007).

*Dimensionality of the omission indicators.* In the latent model-based approach accounting for omitted items, a measurement model for the omission indicators was postulated. It is implicitly assumed that a unidimensional measurement model fits the missing indicators, that is, that the missing indicators measure a single construct representing the tendency for omission. In order to test for this assumption, we applied a Rasch model to the missing indicators of omitted responses and evaluated different item fit measures: weighted mean square (WMNSQ), item characteristic curve (ICC), and point-biserial correlation of the correct response with the estimated ability. This was done separately for missing indicators of omitted items in the reading and mathematical competence test.

*Investigating the amount of nonignorability.* The model-based approaches for nonignorable missing data may be needed if the missing propensity is related to ability. We investigated the relationship of the missing propensities with ability in different models. The models differed in whether the missing propensity was modeled manifest or latent. We estimated the correlation of ability with the latent propensity in the latent model-based approach while we computed the manifest correlation from the regression estimate in the manifest model-based approach. We estimated these correlations for omission tendency, for speed (not-reached items), and using the composite of both kinds of missing responses.

*Performance of the different approaches.* In line with the scaling procedure of the competence tests in the NEPS (Pohl & Carstensen, 2012a), we modeled the observed item responses via a Rasch model (Rasch, 1960). We applied three classical approaches and the two model-based approaches to the data and estimated item difficulty as well

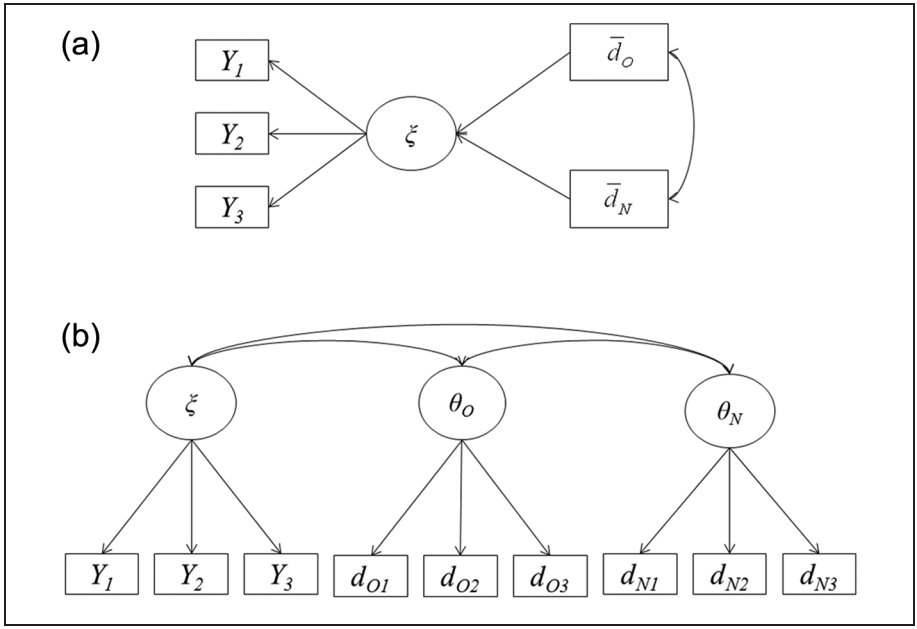


**Figure 2.** Models accounting for omitted (O) and not-reached (N) items simultaneously in one model (a) using manifest missing scores in a multiple latent regression or (b) modeling latent missing propensities in a multidimensional item response theory model.

as person ability. The classical approaches included scoring missing responses as incorrect (M1), ignoring missing responses for the estimation of item parameters and scoring them as incorrect for the estimation of person parameters (M2), and ignoring missing responses in the estimation (M3). Model-based approaches included the manifest approach (M4) and the latent approach (M5). In the latent approach, we modeled the missing indicators using a Rasch model. Note that we used a different latent model depending on the kind of missing response in consideration. For omitted items, we implemented the latent approach of Holman and Glas (2005) while for missing responses due to not-reached items, we used the latent approach of Glas and Pimentel (2008).

We considered both kinds of missing responses, those due to omitted items and those due to not-reached items, in the models in different ways. We performed the analyses separately for omitted items, not-reached items, the composite of omitted and not-reached items, and omitted and not-reached items included separately in one model. Note that when we applied a certain approach to only one kind of missing response, we ignored the other kind of missing response in the analyses. This was because in previous studies, ignoring missing responses proved to provide unbiased results. As a consequence, approach M3 (ignoring missing responses) was the same for all four ways of incorporating different kinds of missing responses.

To evaluate whether the results depend on the content of the test or whether they may be generalized, we performed all analyses for two competence domains, namely reading competence and mathematical competence. All three factors—approach, kind of missing response, and competence domain—were fully crossed. Thus, a 5 × 4 × 2 design resulted.

*Complete case simulation.* The results on the performance of the different approaches in real data analyses give a unique insight into the complex structure of real data sets and in the suitability of the different approaches for empirical data. However, in real data analyses we do not know the true parameters we aim to estimate. In order to corroborate the empirical findings, we conducted a complete case simulation study. For that, we concentrated on the $N = 1{,}143$ subjects in the data set with no missing values in the reading test.[3] We fitted the Rasch model to the 56 reading items and estimated item and person parameters in ConQuest (Wu et al., 2007). In a further step, we systematically introduced missing values due to omission and time limits and evaluated how the different approaches for dealing with missing values may recover the item and person parameters of the complete case analysis.

According to the results found in the empirical analyses, one mechanism of missingness due to omission and two mechanisms of missingness due to time limits were simulated. All of them are based on the results of the empirical analyses. For omission, a latent missing propensity was generated. In line with the results of previous studies (e.g., Holman & Glas, 2005; Rose et al., 2010) and the results of our empirical analyses, the missing propensity was simulated to be normally distributed with a correlation of −.18 with the latent ability. Assuming a unidimensional measurement model of the omission indicators, the difficulty of the omission indicators for each item was simulated to be correlated to the difficulty of the response variables (correlation = −.17). Without introducing any missing values due to not-reached items, a mean omission rate per item of 3.7% ($SD$ = 1.8%) of persons occurred. In this approach, it is assumed that a unidimensional missing propensity exists.

For introducing not-reached items, two mechanisms were simulated that differ in their correlation with the latent ability. This is in line with the empirical findings in this study in which the missing propensity of not-reached items is differently correlated with ability depending on the competence domain considered. In the first condition (not reached with positive correlation), the number of not-reached items was simulated to be positively correlated with ability while in the second condition (not reached with negative correlation), it was simulated to be negatively correlated with ability. Then, the number of not-reached items was truncated, so that all negative values were set to zero. Thus, a skew distributed variable for the number of not-reached items resulted that correlated to .214 and −.277, respectively with reading ability. The mean percentage of not-reached items per persons was 13.4% ($SD$ = 14.1%) and 12.5% ($SD$ = 13.5%), respectively.

The missing generation for omissions was combined with both missing conditions for not-reached items, respectively, so that two simulated data sets resulted, one with

latent omission and not reached with positive correlation and one with latent omission and not reached with negative correlation. All the approaches for dealing with missing values that were applied in the empirical study were then applied to the generated data sets. For the simulation, we only used the most sophisticated approaches of simultaneously accounting for omitted and not-reached items separately in one model. The performance of the different approaches was then evaluated by comparing the estimated item and person parameters from these analyses with those from the complete case analysis. Generation of missing responses and summarizing the results was done in R (R Development Core Team, 2012) whereas IRT analyses were performed in ConQuest (Wu et al., 2007).

## Results

### *Dimensionality of the Omission Indicators*

Since the last item of a test may not contain omitted items (missing responses on the last item are coded as not reached), we only included 58 missing indicators for omitted items in the measurement model for the missing propensity. The missing indicators for omitted responses in the reading test fit a Rasch model quite well. The WMNSQ ranged from 0.84 to 1.28. Four items had a WMNSQ greater than 1.2. Fifty-five indicators had an empirical ICC that well resembled the model-implied ICC. Only for three indicators did the empirical ICC deviate to some extent from the model-implied ICC. These three items were subtasks (dichotomous response format) of the first complex MC item in the test. Of these complex MC items, for all subtasks for which ''false statement'' was the correct response, a considerably high number of missing responses occurred. This may have been an indication of instruction problems. Participants probably ticked only the responses that represented true statements and did not tick an answer when the statement was false. We therefore excluded these missing indicators from further analyses. The point-biserial correlation of the occurrence of missing values on an item and the overall tendency to omit items ranged from .16 to .61 with a mean of .39. The small point-biserial correlations may have resulted from the small number of subjects who omitted an item.

The missing indicators for omitted responses in the mathematics test were also consistent with the Rasch model. The values of the WMNSQ of the items varied between 0.86 and 1.11. The empirical ICCs resembled the model-implied curves very well. The point-biserial correlations of the mathematics items ranged from .16 to .60 with a mean of .38. Overall, besides instruction problems in the reading test, for both competence domains, the omission indicators well fitted a unidimensional Rasch model.

### *Amount of Ignorability*

Investigating the amount of nonignorability, the relationship between the missing propensities and ability was different for different competence domains. In reading, there

was a negative correlation for average omission rate (correlation[manifest] = −.203, $p < .05$) and a positive correlation for the average rate of not-reached items (correlation[manifest] = .102, $p < .05$). Participants with a low-ability estimate tended to omit more items than persons with a high-ability estimate in reading. In contrast to omission, students with a high-ability estimate were those with the larger number of not-reached items. This may indicate that participants who work very carefully on answering the test items are more likely to solve the items correctly but do not reach the end of the test. There was a small positive correlation of omission rate and difficulty of an item (correlation = .165, $p = .23$). The more difficult an item was, the more likely it was to be omitted.

The missing process was different for not-reached items. Since the items in the test were not sorted by difficulty but rather unsystematically distributed in the test, there was no correlation between the missing rate for not-reached items and the difficulty of an item. Modeling latent instead of manifest missing propensities resulted in a similar correlation for omitted items (correlation[latent] = −.175) and a higher correlation for not-reached items (correlation[latent] = .200). Thus, the manifest missing propensity is affected by measurement error that blurs the correlation between ability and missing propensity. Since omission rate and the tendency not to reach the end of the test were differently related to ability, using a composite of the different missing types distorted the correlation of missing propensity and ability. The correlations for the composite score were −.028 and .001 for manifest and latent missing propensity, respectively.

The correlations were different for mathematical competence. Modeling manifest missing propensities, no substantial correlation could be found for any of the missing types (correlation[manifest] = −.002 for omitted items, correlation[manifest] = −.004 for not-reached items, and correlation[manifest] = −.004 for the composite of omitted and not-reached items). Modeling latent instead of manifest missing propensities substantively increased the correlations to −.205 ($p < .05$) for omitted items, −.039 ($p < .05$) for not-reached items, and −.214 ($p < .05$) for the composite of omitted and not-reached items. Thus, using a latent missing propensity accounts for fallible measures and provides a better estimate for nonignorability. In contrast to reading competence, both missing propensities correlated positive with ability. Highly able students in mathematics had lower rates of omitted and not-reached items. This may also have been a result of less speediness in mathematics. Unlike reading, most students were able to reach the end of the test.

The correlations clearly showed that for both kinds of missing responses, missing propensity and ability are not distinct and, thus, that the missing mechanism is not ignorable. Thus, it seemed advisable to account for nonignorability and to include the missing propensity in the model. The relationship between missing propensity and ability was different for omitted and not-reached items, indicating a different underlying missing mechanism. Furthermore, modeling a manifest instead of a latent missing propensity considerably lowered the correlation. Using a composite score across the

different types of missing responses did distort the correlation. Thus, the two kinds of missing responses should be considered separately.

## Performance of the Different Approaches

*Item parameter estimates.* The consistency of the item parameter estimates of the different approaches for reading is depicted in Figure 3. The results of mathematics were similar, however, due to the lower number of missing responses being far less pronounced. Considering only omitted items, besides the approach that scores missing responses as incorrect for item parameter estimation (M1), the item parameter estimates of all other approaches correlated perfectly (correlation = 1.0), indicating a perfect consistency of the estimates. In reading, the item parameter estimates of approach M1 deviated from the estimates of the other four approaches (correlation = .96). Here, the items that had a high omission rate in particular had a different difficulty estimate in M1 as compared with the other approaches. In mathematics, where the number of omitted items was slightly lower, there was no difference in item parameter estimates between any of the approaches (correlation = 1.0).

We found similar results for dealing with only not-reached items[4] for the composite of omitted and not-reached items, and for the separate modeling. There was no difference in item parameter estimates between the last four approaches (M2 to M5, correlation = 1.0). However, the parameter estimates from M1 differed from the other four approaches (correlation = .88 for not-reached items and correlation = .86 for composite score and separate scaling). Again, since there were hardly any not-reached items in mathematics, for mathematics items, item parameter estimates were very similar across the five approaches (correlation = 1.0 for not-reached items and correlation = .99 for composite and separate scaling).

Summarizing the results, scoring missing responses as incorrect (M1) resulted in different difficulty estimates than in approaches where the missing responses were ignored, and there was no difference in item parameter estimates between ignoring missing responses in item parameter estimation and using model-based approaches. This result was consistent for all types of missing responses (omitted, not reached, composite, and separate modeling). In contrast to reading competence, the results for mathematical competence were less striking since the number of missing responses was considerably smaller for mathematics than for reading competence. Regarding item parameter estimation, there was hardly any difference in estimates between methods for mathematics. While for a large number of missing responses, item parameter estimates were considerably affected by the chosen approach, with a number of missing responses as was present in the mathematics test, the choice of method had no strong effect on item parameter estimation.

*Person parameter estimates.* Bayesian expected a posteriori (EAP) estimates were used for person parameter estimation. The results of the comparison of the person parameter estimates between the different approaches are depicted in Figure 4 for reading competence. They were similar for treating only omitted items, only not-

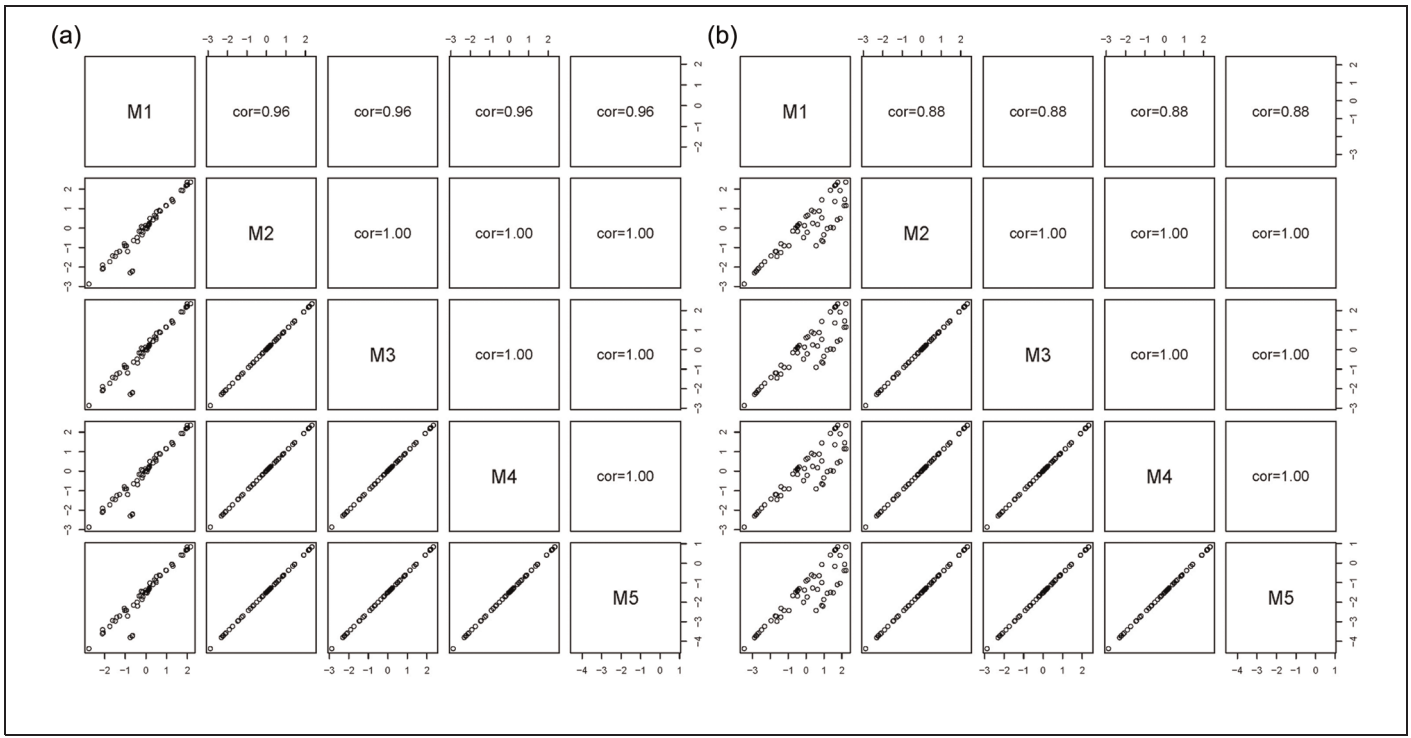


**Figure 3.** *(continued)*

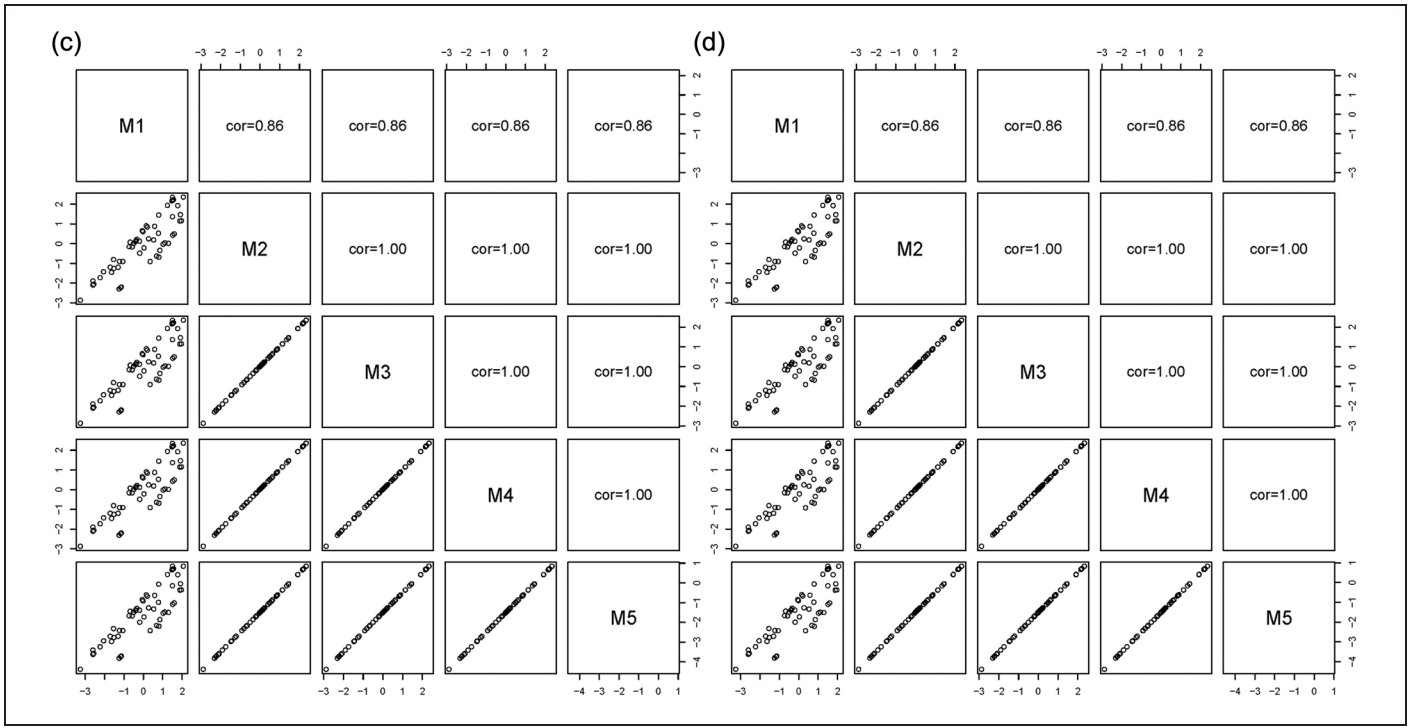


**Figure 3.** Reading competence: Item parameter estimates of different approaches dealing (a) with omitted items only, (b) with not-reached items only, (c) with the composite of omitted and not-reached items, and (d) separately with both kinds of missing responses in one model. Results are shown for scoring missing responses as incorrect (M1), ignoring missing responses for the estimation of item parameters and scoring them as incorrect for the estimation of person parameters (M2), ignoring missing responses in the estimation (M3), manifest model–based approach (M4), and the latent model–based approach (M5).

reached items, the composite of both kinds of missing responses, and for separately accounting for the different kinds of missing responses in one model as well as for the different competence domains. There were two groups of approaches that showed consistent ability estimates within their groups but different ones across groups. The first group was formed by approach M1 and M2, both of which treat missing responses as incorrect when estimating person parameters. The second group was formed by the approach ignoring missing responses (M3) and the two model-based approaches (M4 and M5). Within the groups of approaches, the person parameter estimates correlated to 1.0. Considering only omitted items, across the groups, the person parameters correlated to .94. for reading and between .97 and .98 for mathematics. The results were similar for treating only not-reached items, both kinds of missing responses together, and both kinds separately. The respective correlations between the two groups of approaches varied between .74 and .76 for reading and between .96 and .98 for mathematics. The more missing responses there were, the more the different groups of approaches deviated from each other in their person parameter estimates.

Corresponding to previous research, the approaches scoring missing responses as incorrect resulted in different parameter estimates than approaches ignoring missing responses or model-based approaches. Lord (1974) as well as Mislevy and Wu (1988) analytically derived that scoring missing responses as incorrect results in biased parameter estimates. These results were confirmed in simulation studies (e.g., Rose et al., 2010). The results of the present study showed that scoring missing responses as incorrect considerably changed the estimated person parameters. There was no considerable difference in parameter estimates between ignoring missing responses and any of the model-based approaches. The IRT model ignoring missing responses seems to be robust to violations of the ignorability assumption.

## *Complete Case Simulation*

The results found in the empirical analyses were corroborated in the complete case simulation. Item parameter estimates of the approaches ignoring missing responses and the two model-based approaches correlated across all conditions with the estimates of the complete case analysis to more than .98. The approach of scoring missing responses as wrong (M1 and M2) for item parameter estimation resulted in correlations ranging between .83 and .86, with the estimates of the complete case analysis. This is in line with the empirical findings and suggests that in the simulated conditions, ignoring missing responses and either of the model-based approaches results in unbiased item parameter estimates, while scoring missing responses as wrong results in biased item parameter estimates.

The Bayesian EAP estimates for the complete case analysis as well as for each of the five approaches of dealing with missing values (M1 to M5) is depicted in Figure 5 for both simulation conditions. The person parameter estimates in the simulation corroborate the findings of the empirical studies. There are two groups of

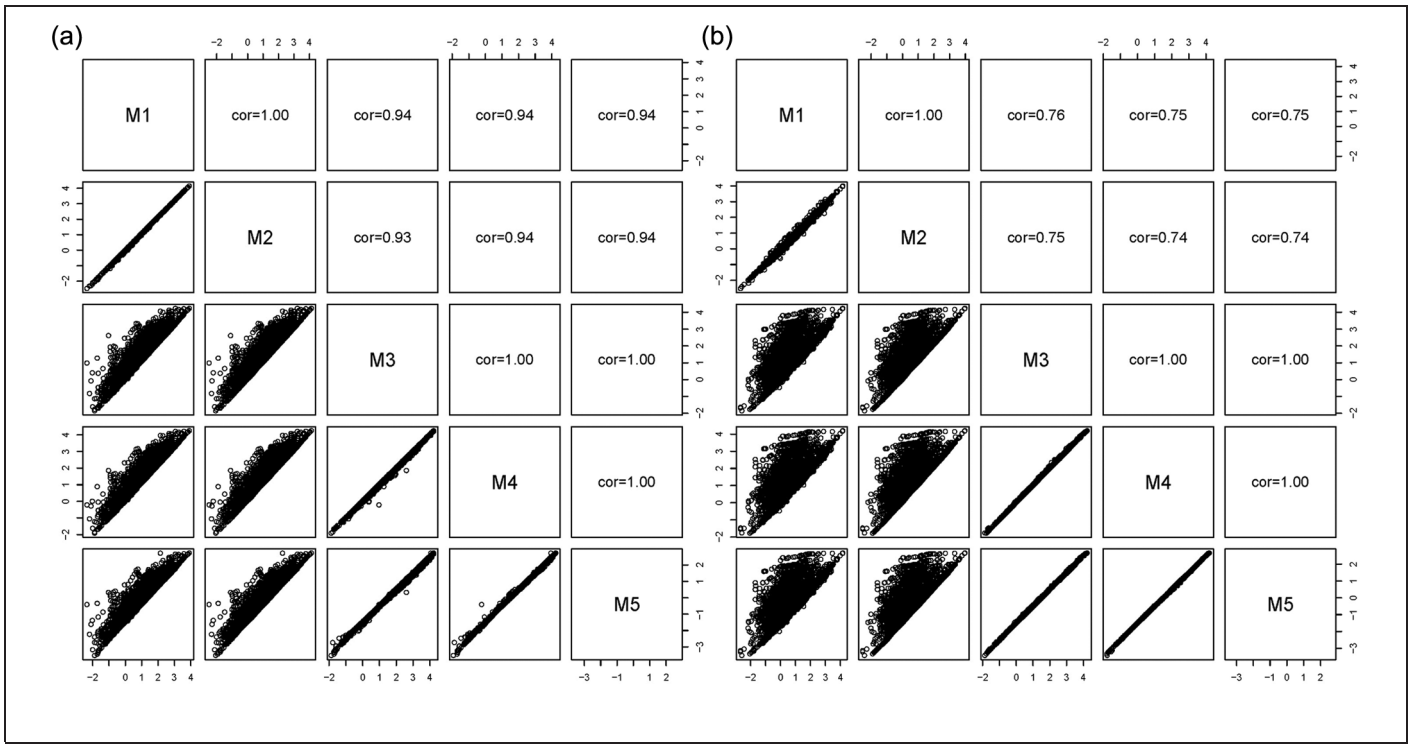


**Figure 4.** *(continued)*

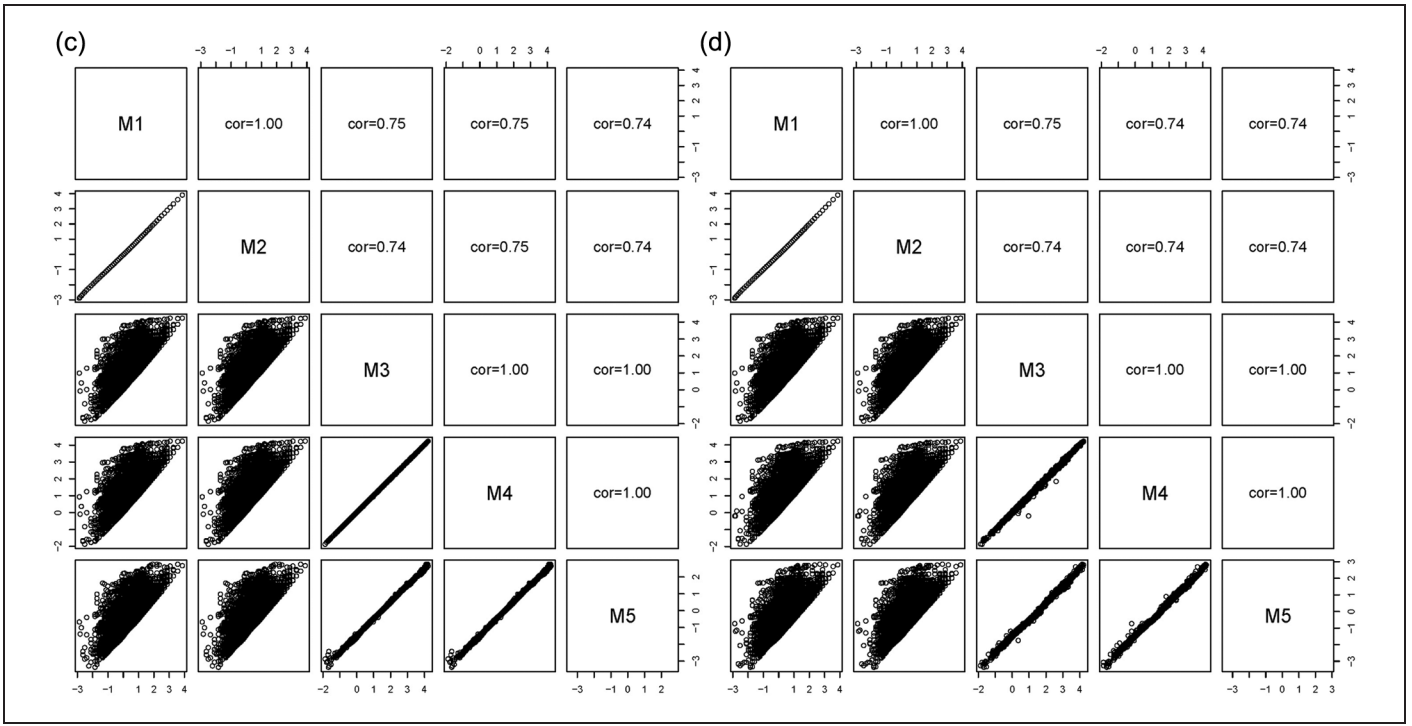


**Figure 4.** Reading competence: Person parameter estimates (expected a posteriori) of different approaches dealing (a) with omitted items only, (b) with not-reached items only, (c) with the composite of omitted and not-reached items, and (d) separately, with both kinds of missing responses in one model. Results are shown for scoring missing responses as incorrect (M1), ignoring missing responses for the estimation of item parameters and scoring them as incorrect for the estimation of person parameters (M2), ignoring missing responses in the estimation (M3), manifest model–based approach (M4), and the latent model–based approach (M5).

approaches, those that score missing responses as wrong (M1 and M2) and those that ignore missing responses or include the missing propensity additionally in the model (M3 to M5). Within these approaches, the person parameter estimates correlate perfectly while the estimates differ between the two groups of approaches. As the missing mechanism in the simulation was generated based on empirical results of the reading and mathematics data, the correlation between the EAPs of the different approaches ranges between those found in reading and mathematics (between .74 and .98 between the two groups of approaches). Again, this may indicate that the simulation conditions well-depict the empirical data. Furthermore, in the simulation study, we may compare the parameter estimates from the different approaches with the estimates from the complete data set without any missing values. This allows for a direct evaluation of the different approaches. The results show that the person parameter estimates when ignoring missing responses as well as when using model-based approaches correlate very high with those from the complete cases. Although, the correlation is not perfect, these approaches very closely resemble the person parameters. In contrast, the EAP estimates when scoring missing responses as incorrect (M1 and M2) deviate considerably from the estimates in the complete case analysis, indicating biasedness.

## Discussion

We probed the performance of different approaches accounting for missing responses in competence tests using empirical as well as simulated data. Contrary to previous work, we considered both kinds of missing responses—those due to omitted and those due to not-reached items—in one model and tried to find appropriate models accounting for both of them. We specifically incorporated the recently developed model-based approaches, tested the assumptions made in these models, and used them for probing the assumption of ignorability. The results regarding the appropriateness of the model assumptions showed that the omission indicators of the reading and the mathematics test well-fitted a unidimensional Rasch model. Exceptions were three missing indicators in the reading test, which did not fit a unidimensional Rasch model. These three items were the first three complex MC items in the test and missing values on these items obviously occurred due to problems in understanding the instruction. Model-based approaches draw on nonignorability of the missing responses. Nonignorability of missing responses is given when the missingness is correlated with ability. For both competence domains, there were substantial correlations between the missing propensity and ability. While subjects with a low ability in reading and mathematics tended to omit more items than highly able students, the relationship between the not-reached missing propensity and ability was different for both competence domains. Students with a high-ability estimate in reading tended to not reach more items than students with a low-ability estimate. This correlation was reversed for mathematics. The students with a low mathematical ability tended to have more not-reached items. This may also have been an effect of the low number

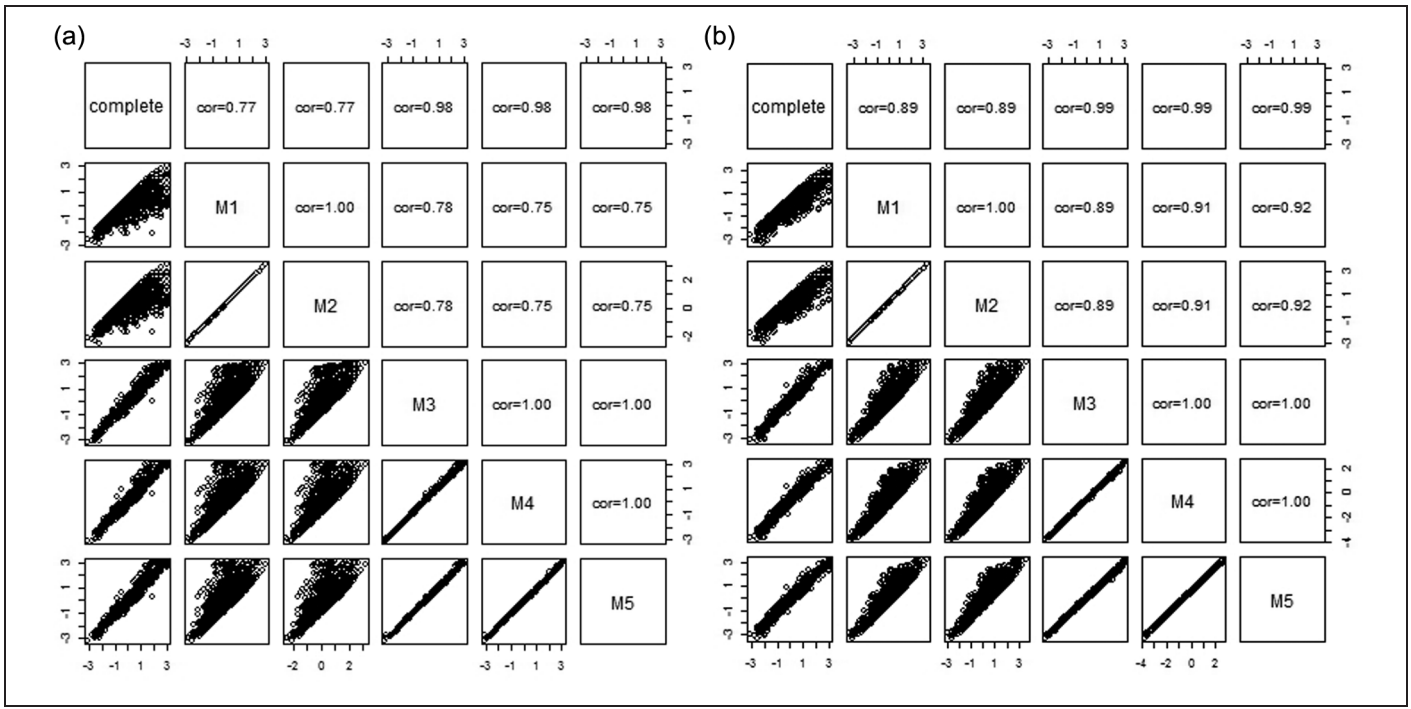


**Figure 5.** Results of the complete case simulation for reading: Person parameter estimates (expected a posteriori) of the different approaches on data with (a) latent omission and not reached with positive correlation and (b) latent omission and not reached with negative correlation. Results are shown for the complete data set, scoring missing responses as incorrect (M1), ignoring missing responses for the estimation of item parameters and scoring them as incorrect for the estimation of person parameters (M2), ignoring missing responses in the estimation (M3), manifest model–based approach (M4), and the latent model–based approach (M5).

of not-reached items in mathematics in general. The correlations increased when modeling a latent instead of a manifest missing propensity. Thus, a latent modeling of the missing propensity may be more effective in accounting for nonignorable missing responses than a manifest approach. As the correlation of missing propensity and ability was rather different for different kinds of missing responses, a composite score of both missing types distorted the correlation. This was especially prevalent in the reading test. When the missing propensity is used to model the item responses, it should be refrained from using composite scores of both kinds of missing responses.

We investigated the performance of different approaches by comparing item and person parameter estimates between approaches. Analytic derivations (e.g., Lord, 1974) and simulation studies (e.g., Finch, 2008; Rose et al., 2010) have shown that scoring missing responses as incorrect results in biased parameter estimates. In line with the existing literature (e.g., Finch, 2008; Lord, 1974; Rose et al., 2010), in the present study, we found different item and person parameter estimates for scoring missing responses as incorrect than for ignoring missing responses. We found no differences in parameter estimation for ignoring missing responses or any of the model-based approaches. Although there was a nonnegligible correlation between ability and missing propensity, and missing responses were, thus, not ignorable, the missing propensity was not needed to model item responses. The IRT model seemed to be quite robust to violations of the ignorability assumption for missing responses in empirical settings like the one we investigated here. This was corroborated in the complete case simulation where we simulated different missing mechanisms. This is in line with findings from Hoshino (2005), who concluded that the observed item responses (or even just the latent variables in the measurement model) are sufficient to account for missing responses in competence tests. This may be due to the fact that in a well-designed test instrument, item responses are highly correlated and the missing responses may well be predicted from the observed responses. However, ignoring missing responses in the estimation may encourage test takers to omit items as a test-taking strategy. This approach may therefore be inappropriate for high-stakes assessments, and model-based approaches may be a good alternative.

Although, competence testing in NEPS is similar to that of other large-scale assessments, the results found for NEPS do not necessarily need to generalize to other studies. Especially in high-stakes assessments, the missing mechanism may be different and may, thus, differently need to be accounted for. There are indications that the results we found in this study do generalize to low-stakes assessments within that age group. In analyses on mathematics and reading competence in Grade 9, we found similar results than those presented for Grade 5 in this article. Ignoring missing data and model-based approaches for nonignorable missing data resulted in similar results, while scoring missing data as incorrect resulted in different parameter estimates. The difference in parameter estimates between these two groups of approaches depended on the amount of missing data. Rose et al. (2010) found similar results for mathematics, reading, and science competence in PISA 2006. However, as the test-taking strategies may differ between low- and high-stakes assessments as

well as between different age groups (e.g., students in primary school, university students, and adults), these results may not generalize to all studies. The analyses conducted here may then well be used in other studies to investigate ignorability of different kinds of missing data and to evaluate whether nonignorable missing data explicitly need to be accounted for using model-based approaches.

As could be shown, IRT-based item and person parameter estimates turned out to be fairly robust to violation of the MAR assumption (Rose, 2013). However, model-based approaches do not only reduce bias but can increase accuracy of parameter estimates. For example, even if the MAR assumption holds true, missingness can be correlated with the latent ability. Using model-based approaches reduces the shrinkage effect and standard errors of EAP and maximum a posteriori person parameter estimates (Rose, 2013). Recently, Rose (2013), Rose, von Davier, and Nagengast (2013), and Rose and von Davier (2013) introduced further alternative models and stepwise procedures to handle nonignorable missing data due to omitted and not-reached items. Similar to the approaches discussed in this article, these methods allow us to adjust for nonignorable missing data as well as increase the accuracy of item and person parameter estimates. The different model-based approaches differ in the underlying assumptions and the model complexity. Further research and more applications of these models in real assessments are needed to answer the question of suitability and practicability of the proposed methods in educational and psychological assessments.

Although model-based approaches for nonignorable missing data do not seem to be needed accounting for missing responses in competence tests, they are very well suited for investigating the occurrence of missing responses. In our analyses, we investigated the relationship of the missing propensity and ability and found indications of differential missing processes for different kinds of missing responses and different competence domains. While omission rate was clearly negatively related to both reading and mathematical ability, the propensity not to reach the end of the test was different between competence domains. In the reading test, in which substantial problems with testing time occurred and where, as a consequence, subjects did not reach the end of the test, missing propensity was negatively related to ability. This may have been an indication of a test-taking strategy: Students who worked very carefully on the test items were more likely to solve an item correctly and consequently, got a higher ability estimate and had more missing responses due to not-reached items. This was less prevalent in the mathematics test, since this test suffered much less from speediness. In further studies, the occurrence and the process leading to missing responses should be investigated in more detail. For understanding the process of missingness, it would surely be valuable to investigate the stability of the missing propensity across competence domains and time (Pohl & Carstensen, 2012b, 2013). Is there a general missing propensity that is some kind of a personality variable or is the missing propensity test-specific? Is the missing propensity stable across time and therefore some kind of a trait, or is it state-specific? Results of such analyses will provide further information on the occurrence of missing responses.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by third-party funds from the German Federal Ministry of Education and Research (BMBF). This research used data from the National Educational Panel Study (NEPS). The NEPS data collection is 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.

## Notes

1. We also conducted the analyses aggregating the subtasks of the complex multiple-choice and matching items to polytomous scores and using a Partial Credit Model for analyses (see Pohl & Carstensen, 2012a for the scaling procedures in National Educational Panel Study). The results focused on in this article treating the subtasks as single items do not differ from using polytomous scores.
2. As a result of data-processing issues, the number of students used in the analyses for assessing reading competence differed slightly from the number used for assessing mathematical competence.
3. We used the data from the reading test because for this test, the largest number of missing values occurred.
4. Modeling a latent missing propensity due to not-reached items while imposing a linear assumption on the item difficulties of the missing indicators resulted in convergence problems. As a consequence, we relaxed the linearity assumptions and allowed for a quadratic relationship between the difficulty of the missing indicators and the position of the item in the test. With such a relaxed constraint, the model converged.

## References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report*. Washington, DC: National Center for Education Statistics.

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.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process—The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. In H. P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* [Special issue] (*Zeitschrift für Erziehungswissenschaft, 14*). Wiesbaden: VS Verlag für Sozialwissenschaften.

Cai, L. (2010). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307-335.

Campbell, D. T., & Erlebacher, A. (1975). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 1, pp. 597-617). Beverly Hills, CA: Sage.

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

De Ayyala, 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.

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, Germany: 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.

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

Harel, O., & Schafer, J. (2003). Multiple imputation in two stages. In *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*.

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.

Hoshino, T. (2005). A latent variable model with non-ignorable missing data. *Behaviormetrika, 32*, 71-93.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. von Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221-244). New York, NY: Springer.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

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: University of California.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 531-543.

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

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48*, 477-482.

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*, 615-630.

Macaskill, G., Adams, R. J., & Wu, M. L. (1998). Scaling methodology and procedures for the mathematics and science competence, advanced mathematics and physics scale. In M. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study. Technical*

*report: Vol. 3. Implementation and analysis* (pp. 91-120). Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Martin, M. O., Mullis, I. V.S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

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. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J. (1993). Should ''multiple imputations'' be treated as ''multiple indicators''? *Psychometrika, 58*, 79-85.

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Rep. no. RR-88-48-ONR). 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* (ETS Research Rep. no. RR-98-30-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, 163*, 445-459.

Mullis, I. V.S., Martin, M. O., & Diaconu, D. (2004). Item analysis and review. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 225-252). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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.

Organisation for Economic Co-operation and Development. (2009). *Pisa 2006 technical report*. Paris, France: Author.

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

Pohl, S., & Carstensen, C. H. (2012b). *Modeling missing data in competence tests*. Project founded in the Priority Programme 1646 ''Education as a lifelong process'' of the German Research Foundation (DFG).

Pohl, S., & Carstensen, C. (2013). Scaling the competence tests in the National Educational Panel Study—Many questions, some answers, and further challenges. *Journal of Educational Research Online,* 5, 189-216.

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, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

R Development Core Team (2012). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputations. *Journal of the American Statistical Association, 102*, 1462-1471.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Unpublished doctoral dissertation). Friedrich-Schiller-University of Jena, Germany.

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 Rep. no. RR-10-11), Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976) Inference and missing data. *Biometrika, 63*, 581-592.

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

Shen, Z. (2000). *Nested multiple imputation* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

Sheriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology, 45*, 81-90.

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.

Stocking, M. L., Eignor, D., & Cook, L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true score equation procedures* (ETS Research Rep. no. RR-88-41). Princeton, NJ: Educational Testing Service.

van den Wollenberg, A. L. (1979). *The Rasch model and time-limit tests* (Unpublished doctoral dissertation). University of Nijmegen, Netherlands.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* [Special issue] (*Zeitschrift für Erziehungswissenschaft, 14*). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Wu, M., Adams, R., & Haldane, S. (2007). ConQuest [computer software]. Melbourne, Victoria: Australian Council for Educational Research.