Article

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

DOI: 10.1177/0013164414561785
epm.sagepub.com

Taking the Missing Propensity Into Account When Estimating Competence Scores: Evaluation of Item Response Theory Models for Nonignorable Omissions

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

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 nonignorable 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 with a model ignoring

[1]Otto-Friedrich University Bamberg, Bamberg, Germany
[2]Free University of Berlin, Berlin, Germany

Corresponding Author:
Carmen Köhler, Otto-Friedrich University Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany.
Email: carmen.koehler@uni-bamberg.de

missing responses. Implications for ability estimation in large-scale assessments are discussed.

Keywords

missing data, nonnormal 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). To enable educational monitoring, data on student knowledge are 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 nonignorable (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 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 National Assessment of Educational Progress (NAEP) 1990 study in Grade 12 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 algorithm (Dempster, Laird, & Rubin, 1977), or multiple imputation (Rubin, 1987).

Many studies have investigated the performance of the aforementioned methods, illustrating their strengths and limitations. In 1974, Lord 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; Hohensinn & 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 expectation-maximization algorithm, the multiple imputation 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 needs to be investigated separately for each missing type. When items are missing due to the design, the researcher can control for the process that 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 article focuses on missing responses that are due to omissions and draws on a model-based approach developed by O'Muircheartaigh and Moustaki (1999), later extended by Holman and Glas (2005). This particular approach tries to take nonignorable 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 **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 **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 **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 **X**, $d_{iv} = 0$. Note that higher missing propensity values indicate less missing responses. To make inferences on examinee proficiency while accounting for nonignorable nonresponse, 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

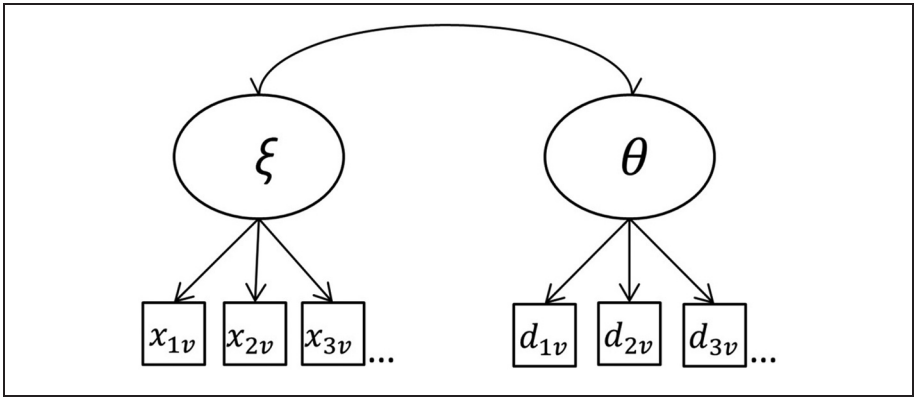

Figure 1. Between-item-multidimensional Item Response Theory model to account for nonignorable omissions.

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 nonignorable omissions.

The marginal maximum likelihood 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 person's 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 $\mathrm{Var}(\xi)$ and $\mathrm{Var}(\theta)$, and the covariance $\mathrm{Cov}(\xi, \theta)$. 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 nonignorable 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 nonignorability and (b) the consequence of using a unidimensional IRT model in which missing responses are ignored. The extent of nonignorability is estimated by the size of the relationship between the missing propensity and the ability. If nonignorability 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 nonignorable omissions with those obtained from the unidimensional IRT model ignoring missing responses. They only found minor differences in ability estimates, even though a nonignorable missing mechanism existed in the data. The violation to ignorability was small, however, and parameters showed robustness to slight violations of ignorability (cf. Holman & 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 that 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, Stancavage, & Reed, 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 nonnormality of the latent distribution when using marginal maximum likelihood 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 den 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 a threat to an application of the model-based approach to actual data.

Because of an incorrect model specification when applying the model-based approach to account for nonignorable 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 nonignorable 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 as follows: 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 nonignorable 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 nonignorable 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 NEPS (Blossfeld, Roßbach, & von Maurice, 2011). One main objective of the NEPS is to collect longitudinal data on competence development (Blossfeld et al., 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 and 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. Because of 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 nonignorable, the model-based approach accounting for nonignorable omissions by Holman and Glas (2005) is needed. We examined the amount of nonignorability 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 unidimensionality 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. 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 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 affects 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 the missing propensity on items with a simple response format; θ_2 represents the 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 $\theta = (\theta_1, \theta_2)$. The MP2D model was applied to the ICT, science, and reading comprehension data. In mathematics, the number of items that 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 considered the Akaike 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 nonignorable omissions (Holman & 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 MP1D is denoted by θ (see Figure 1). Our ABILITY_MP2D model is an extension of this model, in which the missing propensity is modeled 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 model (McCutcheon, 1987; Xu & von Davier, 2008), which departs from the IRT concept that 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

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

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. $P(x_i = k|h)$ 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 six moments, and a structured GDM with a maximum number of 2 moments to describe the distribution. The log-linear model describing the latent skill space $P(h)$, or $P(\xi, \theta)$, 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 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 5. Because of the limited variance of the missing propensity variable, the fewer number of skill levels for the missing propensity sufficed to adequately reflect the skill space. When modeling the latent skill space using 6 moments, 7 additional parameters—4 higher order moments for ability and 3 for missing propensity—were estimated.[3] In sum, the ABILITY_MP1D model was fitted using the 3 distributional alternatives. The convergence criterion was a .0001 minimum change in deviance. 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 that 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 the MP1D and the MP2D model. 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 with 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

Table 1. Unidimensional Missing Propensity (MP1D) and Two-Dimensional Missing Propensity (MP2D) Model Fit Statistics.

Domain (sample)	Model	AIC	BIC	Deviance	LRT	*df*	*p*	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

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; LRT = *likelihood ratio test;* ICT = information and communication technologies.

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.

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 six moments. In the reading data of the adult sample, the BIC was smallest for the model using only two moments.

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

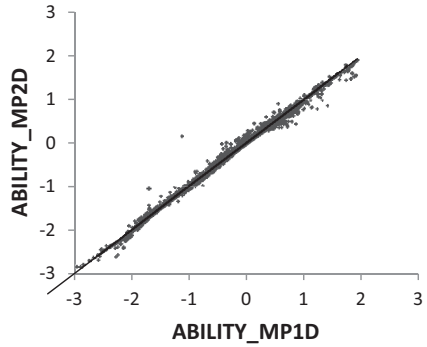

Figure 2. Impact of dimensionality of the missing propensity on ability estimates: Comparing expected a posteriori ability estimates from the model-based approach including a unidimensional missing propensity (ABILITY_MP1D) and the model-based approach including a two-dimensional missing propensity (ABILITY_MP2D).

Impact of distributional assumptions on person parameter estimates. 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 that preserves fewer characteristics of the actually observed joint distribution leads to strongly deviating ability estimates.[4]

For comparisons in all domains and age-groups, the correlations between the EAPs from the saturated and the two-moment models were always smaller than those from the saturated and six-moment models, indicating that the model using six moments better approximated the EAPs from the saturated model than the model using two 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 nonignorable 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

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

Note. ICT = information and communication technologies; AIC = Akaike information criterion; BIC = Bayesian information criterion.

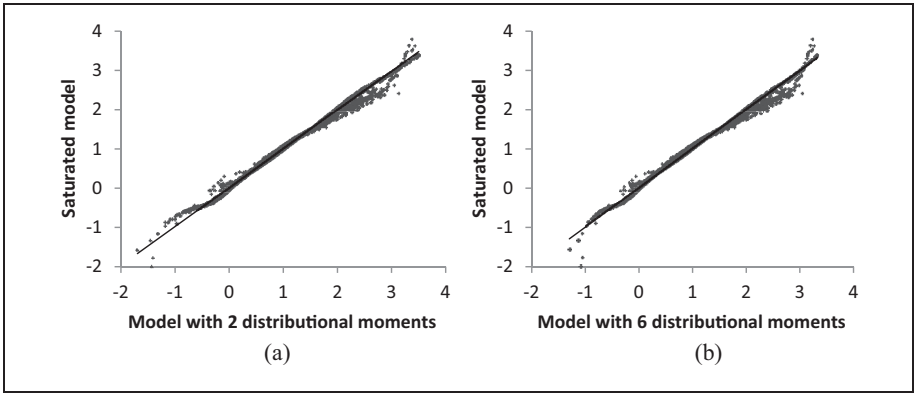

Figure 3. Comparison of ability estimates from the model-based approach including a unidimensional missing propensity (ABILITY_MP1D) for (a) the saturated model and the model using two moments, and (b) the saturated model and the model using six moments.

accounted for or whether the much simpler model in which missing responses are ignored suffices to account for nonignorable 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 with 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.

In contrast to results from previous studies that used stricter distributional assumptions, the missing propensity seems to be needed 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, nonignorable omissions are not accounted for, thus resulting in different person parameter estimates.

Discussion

The present study focused on adequately accounting for nonignorable omissions in competence tests in large-scale assessments. We compared the model by Holman

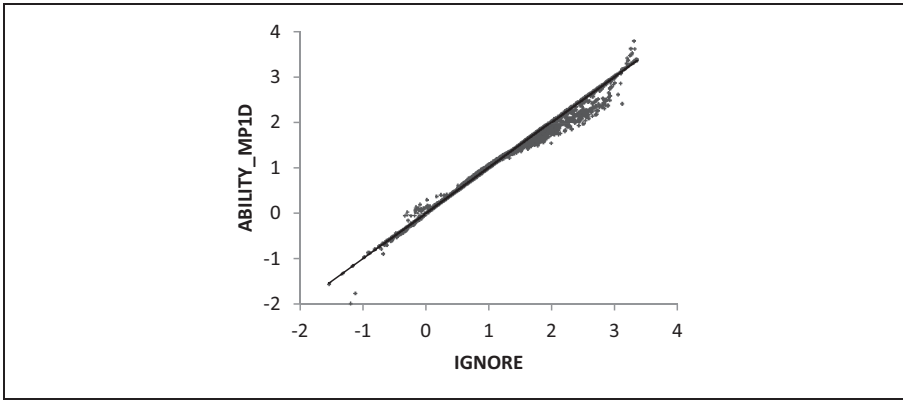

Figure 4. Impact of including the missing propensity in the model: Comparing expected a posteriori (EAP) ability estimates from the unidimensional ability model ignoring missing responses (IGNORE) and the model-based approach including a unidimensional missing propensity (ABILITY_MP1D). In both models,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

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's 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 that 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 subgroups 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 affects 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 nonignorable 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, 2014). The missing values could therefore still be nonignorable with regard to other unobserved variables. In the literature, some models exist which include additional covariates 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 ignorability 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.

Authors' Note

This article 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 article 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).

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

Notes

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 provided by NEPS, items with complex multiple-choice format are not distinguished from matching task items.
3. Although six moments are modeled with regard to the ability dimension and only five moments are modeled with regard to the missing propensity, we will continue referring to this model as the model using six 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.

References

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

Adams, R., & Wu, M. (2007). The mixed-coefficients multinomial logit model: A generalized 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, NY: 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, Hungary: 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). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Government Printing Office.

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.

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

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 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: Statistical Methodology*, *39*, 1-39.

Duchhardt, C., & Gerdes, A. (2012, December): *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. 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, Germany: 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:10.1177/001316440831 5262

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, 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* (Tech. Rep. 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: 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, March). *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

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. (2014, April). *Investigating mechanisms for missing responses in competence tests*. Paper presented at the 2nd Colloquium of the SPP1646 Priority Programme. Florence, Italy.

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. 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*. Newbury Park, 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, Netherlands: 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: Statistics in Society*, *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*, *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: Statistics in Society*, *162*, 177-194. doi: 10.1111/1467-985X.00129

Pohl, S., & Carstensen, C. H. (2012, October). *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, October). *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.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Doctoral dissertation, Friedrich-Schiller-University Jena, Germany). Retrieved from http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-27809/Diss/NormanRose.pdf

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, NY: 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. *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, August). *Data manual: Starting Cohort 6: Adult education and lifelong learning* (Release 3.0.1. NEPS Research Data Paper). Bamberg, Germany: University of Bamberg. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/3-0-1/SC6_3-0-1_DataManual_en.pdf

Skopek, J., Pink, S., & Bela, D. (2013). *Data manual: Starting Cohort 4: Grade 9* (SC4) (NEPS SC3 Version 1.1.0. NEPS Research Data Paper). Bamberg, Germany: National Educational Panel Study, 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* (ETS Technical Report No. 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, Germany: Springer.

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

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model for NAEP data* (ETS Research Report 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