Psychological Test and Assessment Modeling, Volume 57, 2015 (4), 499-522

Investigating mechanisms for missing responses in competence tests

Carmen Köhler[1], Steffi Pohl[2] & Claus H. Carstensen[3]

Abstract

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

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

[1] *Correspondence concerning this article should be addressed to:* Carmen Köhler, PhD, Otto-Friedrich-University Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany; email: carmen.koehler@uni-bamberg.de

[2] Free University Berlin, Germany

[3] Otto-Friedrich-University Bamberg, Germany

Theoretical background

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

Although test developers aim at maximizing the response rates in order to decrease uncertainty regarding the validity of the results, missing values still occur. Most prominent among them are those due to *not-reached* and *omitted* items. The former refer to items towards the end of the competence test, which the examinee did not reach as a result of time limits. The latter are intentionally skipped items within the test. The amount of missing values in competence tests is quite remarkable. In the PISA 2000 study, for example, where one testing session contained about 65 items, the average number of omitted and not-reached items was 2.5 and 1, respectively (Adams & Wu, 2002). These numbers varied considerably between states, ranging from only 0.5 and 0.1 in the Netherlands up to 5 and 4.5 in Brazil. In 2009, the average number of missing items was 5 for omitted items and 2 for not-reached items (OECD, 2012). Here the missing rates were highest for some of the OECD partner countries, with, on average, more than 12 omitted and 2 not reached items in Albania. When looking at the amount of missing values per item in, for example, the 1990 National Assessment of Educational Progress (NAEP) mathematics test, not-reached rates were higher than omission rates. 13 out of the administered 144 items in grade 12 had omission rates above 10%, and not reached rates above 15% (Koretz, Lewis, Skewes-Cox, & Burstein, 1993).

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

ted items are scored as partially correct, with a score corresponding to the reciprocal of the number of options given on a multiple-choice item (Johnson & Allen, 1992).

Each of the aforementioned approaches involves certain assumptions regarding the occurrence of missing responses. Some concerns exist whether these assumptions hold. Treating missing values as incorrect implies that the missing mechanism is purely determined by ability. Furthermore, it presupposes that the participant attempted the item, but could not produce the correct answer. Studies showed, however, that people fail to respond to items for other reasons than lack of knowledge, such as insecurity about the phrasing of the question or lack of motivation (Jakwerth, Stancavage, & Reed, 1999). This is an argument against treating all missing values as if the participant could not have answered them correctly. The approach of scoring missing values as fractionally correct solves the problem of assuming that an examinee performs worse than guessing, but remains deterministic with regard to the value for the missing response (Rose, 2013). The approach of ignoring missing responses implicitly assumes that the missing mechanism is ignorable (Mislevy & Wu, 1988). According to Rubin (1976), the ignorability assumption holds when the missing data are *missing at random* (MAR), and the parameter vector of the probability density function of the missing-data matrix is distinct from the parameter vector of the probability density function of the complete data matrix. These conditions are usually violated in large-scale assessments. The missing mechanism often depends on the unobserved latent ability, and the parameter vectors are thus not distinct from each other (e.g., Glas & Pimentel, 2008; Holman & Glas, 2005). Overall, violations of the assumptions may lead to biased estimates when applying any of the mentioned approaches.

In an attempt to take the non-ignorable missing mechanism into account, researchers have developed models that include the missing mechanism in the measurement model for ability. The idea behind these model-based approaches is that the missing data holds information on the true distribution of the unobserved latent trait, and should thus be incorporated into the model. Most prominent among the approaches are selection models (Heckman, 1976) and pattern mixture models (Glynn, Laird, & Rubin, 1986; Rubin, 1987). They both attempt at modeling the joint distribution of the missing mechanism and the mechanism for the observed responses, and only differ in their specification of this joint distribution. Selection models and pattern mixture models, however, have their limitations in terms of parameter specification and identification, and are rarely applied in practice. O'Muircheartaigh and Moustaki (1999) have developed the approach of modeling the joint distribution further, using multidimensional IRT models. Adaptations of their approach resulted in models for omitted (Holman & Glas, 2005) and models for not-reached items (Glas & Pimentel, 2008), as well as in models that simultaneously account for not-reached and omitted items (Rose, 2013). The great contribution and advantage of these model-based approaches over the previously described approaches is that they consider non-ignorability of the missing data. One challenge for these models lies in finding ways of incorporating the missing mechanism in the measurement model. Analog to efforts regarding an adequate scaling model for persons' abilities, the missing mechanism deserves equal consideration in terms of a proper representation. A first step

towards establishing how the missing process can be modeled involves determining under which circumstances and for what reasons missing values occur.

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

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

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

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

Research questions

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

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

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

Method

Data

The current study employed data from the National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011). In NEPS, a multi-cohort sequence design serves as the basis for data acquisition on competencies, competence development, and its determining factors. There are six starting cohorts: early childhood, kindergarten, grade 5, grade 9, college students, and adults. The competencies measured involve fundamental domains, for example *information and communication technologies* (ICT; Senkbeil, Ihme, & Wittwer, 2013), *science* (SC; Hahn et al., 2013), *mathematics* (MA; Neumann et al., 2013), and *reading competence* (RE; Gehrer, Zimmermann, Artelt, & Weinert, 2013), as well as more general, context-free skills, such as *perceptual speed* or *deductive reasoning*. Besides the assessment of competencies, data on relevant background information associated with competence acquisition and progress are also collected.

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

The NEPS competence domains ICT, science, mathematics, and reading were assessed via item sets covering the respective domain. The items were developed in order to fit the Rasch model (Rasch, 1960) or – in case of polytomously scored items – the partial credit model (PCM; Masters, 1982). The tests in ICT, science, and reading mostly contained items with a simple and some items with a complex multiple-choice response format. A simple multiple-choice item consisted of a single question, and required the examinee to choose the correct answer from several presented response options; a complex multiple-choice item entailed several subtasks (i.e., several questions), each containing two response options. In reading competence, few items were matching tasks, meaning that the examinee was asked to match several headings with corresponding text passages. In mathematics, most items had a simple multiple-choice response format, very few items were complex multiple-choice questions, and some were short constructed response items where the participant was required to insert, for example, the solution to a mathematical problem. In the following, we refer to simple multiple-choice items as

having a simple response format and to complex multiple-choice and matching tasks as having a complex response format. All considered samples were tested in mathematics and reading competence. In the ninth-grade sample, additional data was available in the two domains ICT and science. Table 1 gives an overview of the number of administered as well as the average amount of not-reached and omitted items in each cohort and competence domain. Across all domains and all three cohorts, people omitted, on average, between 1.5% and 8% of the administered items. For not-reached items, the numbers ranged between 1.2% and 10.5%. Not-reached rates were especially high in the reading domain in the fifth-grade and the adult sample. In all cohorts, the amount of omissions was higher in mathematics as compared to the amount of not-reached items, whereas the opposite was the case with regard to the reading domain. Overall, the amount of missing values was not negligible, and might therefore need to be considered in the scaling.

Since one aim of our study was to investigate interindividual differences in the missing propensities, we tried to explain the occurrence of not-reached and omitted items via further competencies collected in NEPS, demographic variables, and personality traits. The competencies included *reading speed*, *perceptual speed*, *deductive reasoning*, *procedural metacognition*, and *declarative metacognition.* The reading speed test consisted of 51 sentences making certain statements, which the participant was asked to rate as either true or false (Zimmermann, Gehrer, Artelt, & Weinert, 2012). The test measuring perceptual speed was time-limited. It comprised 93 items, which required the examinee to match numbers to certain symbols in a correct order. For measuring deductive reasoning, the examinee was presented with 12 matrices items (see Haberkorn & Pohl, 2013), which were developed by Lang and colleagues (Brunner, Lang, & Lüdtke, 2009; Lang, Kamin, Rohr, Stünkel, & Williger, 2012). In all three of the aforementioned tests, the achieved sum score served as the indicator of a participant's skill level. Procedural metacognition was assessed via the examinee's judgment of their own performance in the competence domains (see Lockl, 2013). The number of correctly answered items was

Table 1:
Average amount of missing items per person in IRT-scaled competence tests in three NEPS cohorts

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

subtracted from the number of items the participant estimated as *answered correctly*, and subsequently divided by the number of actual items in the test. The thus calculated percent difference gave information on an examinee's over- or underestimation of their abilities in the respective domain. For declarative metacognition, participants were presented with texts in which certain scenarios concerning school or leisure activities were described (see Lockl, 2012). Several planning, organizing, and resource management strategies were suggested, and the examinee rated them in terms of their usefulness on a four-point rating scale (from 1 = *not useful at all* to 4 = *very useful*). The 69 single evaluations regarding eight different scenarios were compared with expert ratings, and scored as either correct or incorrect. The subsequently calculated mean test score gave information on a person's overall declarative metacognitive skills. The sociodemographic variables we investigated were gender *(female* versus *male*), migration background (*yes* versus *no*), and school type. School type was dummy-coded, so that the three dummy-variables indicated whether a person attended (1) lower secondary school, (2) intermediate secondary school, or (3) comprehensive secondary school; upper secondary school served as the reference group. The considered personality traits involved the five NEO-FFI factors and global self-esteem. The NEPS data provided estimated mean scores for *Neuroticism*, *Extraversion*, *Openness*, *Agreeableness*, and *Conscientiousness* from the BFI-10 short version of the NEO-FFI (Rammstedt & John, 2007). Global self-esteem was available as a sum score, resulting from ratings of ten items tapping the self-esteem construct (e.g., "I feel useless.") on a five-point rating scale (with 1 = *does not apply* to 5 = *applies completely*). The demographic variables as well as the personality traits were assessed in a questionnaire subsequent to the competence testing.

Analyses

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

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

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

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

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

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

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

Person-specificity of the missing propensities

In order to test for person-specificity of the missing propensities, we investigated the stability of not reaching and omitting items across different competence domains. The analyses were conducted in all three age cohorts. Note that for the fifth-graders and the adults, only the relationship between the missing propensities in mathematics and reading could be considered, whereas for the ninth-graders, the relationships between the missing propensities in all four competence domains could be examined. To determine the stability of the tendency to not reach items across competence domains, we computed manifest correlations between the sum scores of not-reached items across different domains; to test the stability of the omission tendency, we estimated latent correlations between the omission tendencies of different domains. As the omission tendencies were modeled two-dimensionally in ICT and reading, the respective between-item-multidimensional IRT model for evaluating the stability of omissions was six-dimensional in grade nine – two dimensions for modeling the omission propensity in ICT and reading, respectively, one dimension in science and mathematics, respectively. In grade five and the adult sample, the models were three-dimensional – one dimension in mathematics and two in reading. High correlations indicate that the missing propensities depend less on the competence domain, but are rather person-specific.

Relations between person characteristics and the missing propensities

Our second research question dealt with explaining interindividual differences between peoples' missing propensities. We therefore analyzed whether competencies, socio-demographics, and personality traits influence the tendency to not reach or omit items in all three cohorts, respectively. We conducted these analyses exemplary in the reading domain, and validated our findings based on mathematics data. Five multiple linear regression models were used to determine the relationship between the explaining variables and the tendency to not reach items; five multiple latent regression models were estimated with regard to the omission tendency. All models included the respective other missing propensity (i.e., the one not focused on) as an explaining variable, since previous studies showed dependencies between omitted and not-reached items.[4] The remaining explaining variables were added in blocks in a consecutive order. Since the strongest relations were found between the missing propensities and ability, the first model involved competencies. Model 2 additionally comprised socio-demographics. We thirdly included personality traits, since they might explain interindividual differences in the missing propensities of the examinees beyond what competencies and demographics already elicit. Model 4 involved interactions between the respective other missing propensity and competencies as well as interactions between the ability in the tested domain and other competencies. Model 5 additionally consisted of interactions between the respective other missing propensity and socio-demographics as well as interactions between the competence in the tested domain and socio-demographics. The interactions were added in order to test whether relationships differ for various subgroups. The competencies included were (a) the ability of the respective domain,[5] (b) reading speed, (c) perceptual speed, (d) deductive reasoning, (e) procedural metacognition, and (f) declarative metacognition. Socio-demographics involved gender, migration background, and school type. The personality traits encompassed global self-esteem and the five NEO-FFI factors. Note that not all variables were available in all three data sets. Since no personality tests were administered to the adults, Model 3 was not estimated in the adult sample.

Due to missing values on some of the explaining variables, we imputed them using the R package MICE (van Buuren & Groothuis-Oudshoorn, 2011). The imputation model encompassed all relevant variables of the regression model, including interaction terms, as well as additional predictors explaining other variables or their missing values. The applied imputation methods were *predictive mean matching* for continuous variables, *logistic regression* for binary variables, and the *ordered logit model* for ordered variables with more than two levels. We used passive imputation in order to preserve the relationships of variables included in interaction terms. We chose 20 iterations, producing a

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

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

single imputed data set. Based on the imputed data set, we estimated the five multiple regression models with the not-reached variable as the dependent variable, and the five multiple latent regression models explaining the omission tendency. Note that in the reading domain, the omission tendency was modeled two-dimensionally.

For all manifest analyses, we used the software R (R Core Team, 2014). All analyses including latent variables were conducted in ConQuest (Wu, Adams, Wilson, & Haldane, 2007).[6]

Results

Person-specificity of the missing propensities

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

Table 2:
Manifest correlations between not-reached tendencies of different domains in three cohorts

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

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

The latent correlations between the omission tendencies in different domains are presented in Table 3. Within the same competence domain, the omission dimensions correlated rather high in reading (between $r = .76$ and $r = .81$) in all three cohorts, while they correlate somewhat lower in ICT ($r = .58$). The different omission dimensions are thus more distinct from each other in ICT than in reading. Between different competencies, they were medium to high, ranging from $r = .24$ to $r = .77$. Consequently, a person's omission tendency remained relatively stable across competence domains. In the ninth-grade sample, omission tendencies in ICT, science and mathematics correlated higher amongst each other than with the omission tendency in reading. Thus, the omission tendency in reading deviated more from omission tendencies in the other domains. When comparing the correlations across the cohorts, correlations between the omission tendency in mathematics and the two omission tendencies in reading were similar in all age cohorts. The amount of person-specificity of the omission propensity seemed to be similar in different age cohorts. As expected, the omission tendencies between reading and ICT correlated higher within the same response format.

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

Table 3:
Latent correlations between omission tendencies of different domains in three cohorts

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

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

Relations between person characteristics and the missing propensities

We subsequently investigated which person characteristics explain the missing propensities in the reading domain. As is evident in Table 4, the most prominent predictors of the tendency to not-reach items across all three cohorts was the first dimension of the omission tendency (D1: omission tendency on items with simple multiple-choice format) and reading speed: Students with more omissions on simple multiple-choice items reached

Table 4:
Standardized regression coefficients of multiple regressions to predict tendency to not reach items in reading

	Model 1			Model 2			Model 3			Model 4			Model 5		
Predictors[a]	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**
Missing propensity															
Omission D1	**-0.29**	**-0.50**	**-0.11**	**-0.29**	**-0.48**	**-0.10**	**-0.29**	**-0.48**		**-0.52**	**-1.13**	-0.10	**-0.44**	**-0.57**	0.02
Omission D2	-0.01	-0.01	**-0.11**	-0.01	-0.01	-0.09	-0.01	-0.01		0.00	0.04	**-0.15**	-0.04	0.06	**-0.26**
Competencies															
Reading competence	**0.17**	0.05	-0.08	**0.20**	0.09	-0.04	**0.20**	0.09		**0.75**	**0.46**	-0.05	**0.42**	**0.26**	-0.06
Reading speed	**-0.32**	**-0.17**	**-0.44**	**-0.31**	**-0.14**	**-0.42**	**-0.31**	**-0.14**		**-0.16**	0.01	**-0.40**	**-0.17**	-0.08	**-0.42**
Procedural metacognition	0.06	0.03	**-0.16**	0.06	0.03	**-0.16**	0.06	0.03		**0.10**	-0.07	-0.03	**0.10**	-0.05	-0.03
Demographics															
Migration background (yes vs. no[b])				0.01	0.03	**0.11**	0.01	0.03		0.01	0.03	**0.11**	-0.01	0.04	**-0.11**
Lower vs. upper secondary school[b]				0.09	**0.18**	0.10	0.09	**0.18**		0.10	**0.17**	0.10	**0.18**	**-0.28**	**0.13**
Comprehensive vs. upper secondary school[b]				0.10	**0.11**	NA	0.09	**0.11**		0.09	0.10	NA	0.03	**-0.26**	NA
Personality															
Interactions between															
- omission and competencies															
- ability and competencies															
Omission D1 x ability										-0.05	-0.04	**0.16**	-0.06	-0.10	**0.12**
Omission D1 x reading speed										**0.34**	**0.43**	0.01	**0.33**	**0.27**	-0.04
Omission D1 x perceptual speed										**0.11**	**0.28**	NA	**0.11**	**0.26**	NA
Omission D2 x deductive reasoning										0.02	**0.14**	NA	0.01	**0.15**	NA
Omission D2 x procedural metacognition										-0.02	-0.02	0.00	-0.01	-0.02	0.00
Omission D2 x declarative metacognition										0.04	-0.06	NA	0.07	-0.04	NA
Ability x deductive reasoning										**-0.12**	-0.08	NA	-0.06	-0.05	NA
Ability x declarative metacognition										**-0.42**	-0.15	NA	**-0.28**	**-0.13**	NA
Interactions between															
- omission and demographics															
- ability and demographics															
Omission D1 x gender													0.01	**0.15**	**-0.13**
Omission D1 x migration background													-0.04	-0.08	**-0.31**
Omission D1 x lower secondary school													-0.02	**-0.41**	-0.05
Omission D1 x comprehensive secondary school													-0.04	**-0.24**	NA
Omission D2 x comprehensive secondary school													0.02	-0.11	0.03
Ability x lower secondary school													**0.18**	0.04	0.01
R^2	.246	.332	.348	.258	.350	.365	.258	.352		.293	.379	.370	.304	.340	.380

Note. Standardized regression coefficients with $\beta > .1$ and $p < .05$ are in boldface. SC3 = starting cohort 3 (fifth-graders); SC4 = starting cohort 4 (ninth-graders); SC6 = starting cohort 6 (adults); D1 = latent omission tendency on simple multiple-choice items; D2 = latent omission tendency on complex multiple choice or matching task items, NA = variable not available in data set. [a]Only variables listed where, in any of the regression models, the standardized regression coefficient was $\beta > .1$ and $p < .05$. [b]Serves as the respective reference group ($x = 0$)

fewer items; students with higher reading speed reached more items.[7] Note that solely the first dimension of the omission tendency served as a relevant predictor, meaning that the two omission tendencies differently relate to the tendency to not reach items. Persons' actual ability in reading was only a meaningful predictor for the tendency to not reach items in fifth grade, where, surprisingly, students reached fewer items when their ability in reading was higher. From the demographic variables, migration background and school type were relevant in some of the cohorts: Controlling for competencies, ninth-grade students in lower secondary school or in comprehensive secondary school reached fewer items than students in upper secondary school; adults without a migration background reached more items than adults with a migration background. This indicates that the groups differ for reasons other than their actual competence level. None of the personality variables we added in Model 3 further explained variance of the tendency to not reach items. Consequently, personality variables have no explanatory value with regard to the missing process. In Models 4 and 5, many of the included interactions were meaningful predictors of the dependent variable, especially interactions between omission and other competencies in Model 4, and between omission and demographic variables in Model 5. The relationship between the tendency to not reach items and the omission tendency was therefore not unanimous across all competence levels (with respect to the competencies we investigated) and across all subgroups (with respect to the demographic variables we investigated). Overall, the models explained a substantial amount of variance. R-squared ranged between $R^2 = .25$ and $R^2 = .38$. Table 4 also reveals the highly homogeneous findings across the three age cohorts. Not only were the same predictors relevant, but also the direction of the relationship was identical.

Tables 5 and 6 illustrate the results of the latent regression of the omission tendency on the explaining variables. Note that the omission tendency was modeled two-dimensionally, which allowed us to investigate the relationship between the explaining variables and the omission behavior on simple multiple choice items (D1; see Table 5) and the omission behavior on items with a more complex response format separately (D2; see Table 6). For both dimensions, reading competence and reading speed mainly determined the tendency to omit items: Higher competence levels in reading and higher reading speed concurred with fewer omissions. An additional important variable regarding the tendency to omit D1 items were not-reached items: A higher tendency to not reach items encompassed a higher tendency to omit multiple-choice items. This was in accordance with the above findings, where the tendency to omit on D1 was relevant for predicting the tendency to not reach items. Regarding the omission of items with a complex response format (D2), deductive reasoning was a significant explaining variable: Students with higher deductive reasoning skills rather responded to D2 items. Furthermore, migration background and school type were relevant predictors in some of the cohorts: People without a migration background as well as higher educated people omitted less D2 items even when controlling for competencies. Except for global self-esteem,

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

Table 5:
Standardized regression coefficients of multiple latent regressions to predict omission tendency on simple multiple-choice items in reading

	Model 1			Model 2			Model 3			Model 4			Model 5		
Predictors[a]	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**
Missing propensity															
Not-reached	**-0.15**	**-0.1**	0.01	**-0.14**	-0.09	-0.06	**-0.13**	-0.09		0.07	-0.03	-0.01	0.04	**-0.17**	0.06
Competencies															
Reading competence	-0.02	0.06	**0.15**	-0.02	0.05	**0.12**	-0.03	0.04		**-0.37**	**-0.59**	**0.30**	**-0.40**	**-0.38**	**0.32**
Reading speed	**0.16**	**0.17**	**0.20**	**0.17**	**0.17**	**0.19**	**0.16**	**0.17**		**0.16**	**0.20**	**0.20**	**0.16**	**0.19**	**0.19**
Deductive reasoning	0.03	-0.01	NA	0.02	-0.03	NA	0.02	-0.03		0.03	-0.02	NA	0.03	-0.02	NA
Demographics															
Migration background (yes vs. no[b])				0.01	-0.06	NA	0.01	-0.06		0.01	-0.06	0.02	0.01	-0.06	0.01
Lower vs. upper secondary school[b]				-0.01	-0.04	NA	0.00	-0.05		-0.02	-0.03	-0.08	-0.02	-0.05	**-0.12**
Personality															
Global self-esteem							0.09	-0.01	NA	**0.10**	-0.01	NA	0.10	0.00	NA
Interactions between *- not-reached and competencies* *- ability and competencies*															
Not-reached x reading speed										0.00	**-0.22**	0.00	0.00	**-0.22**	0.00
Not-reached x deductive reasoning										-0.02	0.00	NA	-0.02	0.01	NA
Not-reached x declarative metacognition										**-0.16**	**0.18**	NA	**-0.16**	**0.20**	NA
Ability x reading speed										-0.03	**0.26**	**-0.16**	-0.02	**0.23**	**-0.18**
Ability x perceptual speed										0.03	-0.03	NA	0.03	-0.03	NA
Ability x deductive reasoning										**0.12**	0.07	NA	**0.12**	0.04	NA
Ability x procedural metacognition										**-0.10**	-0.03	-0.02	**-0.11**	-0.02	-0.02
Ability x declarative metacognition										**0.18**	**0.27**	NA	**0.22**	**0.21**	NA
Interactions between *- not-reached and demographics* *- ability and demographics*															
Ability x lower secondary school													0.01	-0.08	-0.04
R^2	.105	.341	.446	.112	.345	.426	.164	.383		.259	.416	.410	.258	.408	.431

Note. Standardized regression coefficients with $\beta > .1$ and $p < .05$ are in boldface. SC3 = starting cohort 3 (fifth-graders); SC4 = starting cohort 4 (ninth-graders); SC6 = starting cohort 6 (adults); D1 = latent omission tendency on simple multiple-choice items; D2 = latent omission tendency on complex multiple choice or matching task items, NA = variable not available in data set.
[a]Only variables listed where, in any of the regression models, the standardized regression coefficient was $\beta > .1$ and $p < .05$
[b]Serves as the respective reference group ($x = 0$)

Table 6:
Standardized regression coefficients of multiple latent regressions to predict omission tendency on items with a complex response format in reading

	Model 1			Model 2			Model 3			Model 4			Model 5		
Predictors[a]	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**	**SC3**	**SC4**	**SC6**
Missing propensity															
Not-reached	-0.08	-0.04	-0.06	-0.08	-0.03	-0.05	-0.07	-0.03		-0.15	-0.12	**-0.11**	**-0.20**	-0.08	-0.07
Competencies															
Reading competence	0.06	**0.16**	**0.42**	0.04	**0.14**	**0.36**	0.03	**0.14**		-0.10	0.06	**0.39**	0.02	**-0.25**	**0.37**
Reading speed	**0.12**	0.08	**0.19**	**0.12**	0.07	**0.20**	**0.12**	0.08		**0.11**	0.10	**0.20**	**0.11**	**0.11**	**0.20**
Deductive reasoning	**0.21**	**0.14**	NA	**0.20**	**0.13**	NA	**0.20**	**0.12**		**0.17**	0.08	NA	**0.16**	0.09	NA
Demographics															
Migration background (yes vs. no[b])				-0.04	**-0.12**	0.04	-0.04	**-0.12**		-0.04	**-0.11**	0.04	-0.04	**-0.14**	0.03
Lower vs. upper secondary school[b]				-0.07	-0.06	**-0.13**	-0.06	-0.07		-0.07	-0.08	**-0.13**	**-0.12**	-0.05	**-0.19**
Personality															
Global self-esteem							**0.12**	0.03	NA	**0.13**	**0.03**	NA	**0.13**	0.03	NA
Interactions between *- not-reached and competencies* *- ability and competencies*															
Not-reached x reading speed										0.09	-0.07	0.00	0.09	-0.07	0.00
Not-reached x deductive reasoning										**0.11**	**0.21**	NA	**0.11**	**0.24**	NA
Not-reached x declarative metacognition										-0.03	0.03	NA	0.00	0.05	NA
Ability x reading speed										0.02	**0.14**	0.03	0.00	**0.20**	0.00
Ability x perceptual speed										-0.05	**0.15**	NA	-0.05	**0.15**	NA
Ability x deductive reasoning										0.04	-0.09	NA	0.00	-0.02	NA
Ability x procedural metacognition										**-0.15**	-0.07	-0.02	**-0.14**	-0.08	-0.02
Ability x declarative metacognition										0.11	**-0.13**	NA	0.09	-0.07	NA
Interactions between *- not-reached and demographics* *- ability and demographics*															
Ability x lower secondary school													-0.08	**0.13**	-0.03
R^2	.193	.404	.396	.204	.371	.415	.228	.373		.261	.342	.419	.266	.368	.432

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

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

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

the included personality variables in Model 3 had no explanatory value for the tendency to omit items. For fifth-graders, global self-esteem enhanced the response behavior to D2 items, meaning that fifth-graders with higher self-esteem attempted more items with a complex response format. Note that the two omission dimensions were explained by quite different variables; hence, the process leading to an omission on a simple multiple-choice item is quite distinct from the process leading to an omission on an item with a more complex response format. In Model 4, the regression models for both omission dimensions (D1 and D2) showed significant interactions between the tendency to not reach items and competencies as well as between the ability in reading and competencies. This indicates that the bivariate relationships in models one to three between the tendency to omit and the tendency to not reach items as well as between the tendency to omit and reading ability was quite different depending on the skill level in other competencies. This should be considered when modeling persons' tendencies to omit items. In Model 5, which included interactions between the tendency to not reach items and demographics as well as between reading ability and demographics, only the interaction between ability and lower secondary school served as an additionally relevant predictor in the ninth-grade sample. In all three cohorts, the models explained the omission tendency to an extensive amount (D1: $.1 < R^2 < .45$; D2: $.2 < R^2 < .43$).[8] As for the tendency to not reach items, the directions of the relationships were, in general, identical across the different age cohorts.

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

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

tiple-choice items was quite distinct from the tendency to omit on items with a more complex response format.

Discussion

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

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

Overall, our results replicate and enhance previous findings. Several studies indicated that the amount of missing responses depends on the actual ability of a person (e.g., Pohl et al., 2014; Rose et al., 2010; Stocking et al., 1988). Our study demonstrates that ability rather plays a role in the omission mechanism than in the mechanism for not reaching items. People with lower ability levels generally tend to omit more items. In our study, this relationship was more pronounced in the mathematics test than in the reading test. We also found other, more domain-general competencies, which related to the missing propensities. Especially reading speed emerged as a dominant factor, explaining both the tendency to omit and the tendency to not reach items. It is interesting to note that even after controlling for the actual ability in the tested domain, slower readers reach fewer items at the end of the test and also skip more items throughout the test. This was the case for all cohorts and not exclusively in the reading, but also in the mathematics domain. Speed obviously plays a relevant role even in low stakes assessments, and needs to be considered in the stage of test development as well as in the scaling. In confirmation with past research (Ben-Shakhar & Sinai, 1991; Grandy, 1987; Koretz et al., 1993; Zhang, 2013), we detected mixed results regarding a gender effect. The tendency to omit

items was the same for males and females in reading, but not in mathematics. In mathematics, female fifth-graders and adults omitted more items than males, even after controlling for all other competencies. This might be due to gender discrepancies with regard to self-efficacy in mathematics (e.g., Louis & Mistele, 2011; Vermeer, Boekaerts, & Seegers, 2000). Migration background and school type were also relevant predictors in some of the cohorts insofar that people with a migration background and a lower educational level showed higher missing tendencies, even after controlling for all competencies. This indicates that other factors not investigated in the current study might account for differences between these subgroups. People with a migration background and a lower educational level possibly refrain from attempting items with a complex response format because they perceive them as more difficult. The factors possibly explaining differences between the aforementioned subgroups certainly need further investigation. The differences should also be considered in the stage of item calibration in order to avoid systematic disadvantages for certain subgroups. In terms of the personality traits examined in our study, none additionally explained participants' missing propensities. They can therefore be disregarded with respect to the missing data mechanism. Various interactions we investigated were relevant, especially those concerning competencies. They serve as moderators between the two missing propensities as well as between ability and the missing propensities, and might need consideration when modeling the missing data mechanism. Lastly, the omission tendencies seemed fairly distinct from each other, and related to different characteristics. These results clearly indicate that missing values on simple multiple-choice items result from a different mechanism than missing values on items with a more complex response format. Since these omission processes differ, they should be handled separately when modeling the missing data mechanism.

In terms of generalizability of the results, we focused on omissions and not-reached items in a low-stakes assessment. In high-stakes assessments, other test-taking strategies might prevail, thus resulting in different missing data mechanisms. Within the framework of low-stakes assessments, however, we could demonstrate person-specificity of the missing propensities in three cohorts with a wide age range. Furthermore, we examined interindividual differences between persons' missing propensities, and showed that, across different age cohorts and two different test domains, the missing propensities equally relate to other characteristics. These results indicate that the missing propensities might be some sort of a construct inherent in a person. According to Cronbach and Meehl (1955), the process of validation involves various inquiries as well as evidence from different sources. Both the stability over occasions and the uniform relationship to other stable person characteristics meet two of the criteria in the validation procedure (Cronbach & Meehl, 1955). Additional indications would be necessary in order to truly validate the missing propensities as constructs, for example by examining persons' missing propensities across various time points using longitudinal data. This would further verify the stability of persons' missing propensities. Although we were able to identify person characteristics that well predicted the missing propensities, some of the variance between peoples' omission and not-reached tendencies was left unexplained. Future research might consider other possible influences. Motivation, for example, plays a role in the performance on low-stakes tests (Wise & DeMars, 2005), also affecting the amount of omissions (Jakwerth et al., 1999). As the missing propensities were, in part,

specific to the tested domain, it would be valuable to investigate further domain related characteristics, such as the self-concept in the respective domain or the fear of failure.

One strength of this investigation is that we integrated research from previous findings, covering a broad spectrum of aspects potentially relevant for explaining the missing propensities. Thus, we identified factors which remain meaningful even after controlling for all others. We were also able to determine some competencies which moderate the relationship between omitted and not-reached items as well as between the missing propensities and ability, and which should therefore be taken into consideration when accounting for missing values. A further novelty of our study was the separation of the omission tendency based on response format. Most large-scale studies make use of several types of response formats, and need to consider that the missing mechanisms differ accordingly. In light of the scaling of competencies and models which aim at including the missing data mechanism in the measurement model, our results demonstrate which variables are relevant in predicting a missing value. The stability of our results further demonstrates that the missing data mechanism is relatively uniform and may be modelled equally across different domains and cohorts. Including the missing propensity as well as relevant variables in the measurement model for ability might enhance the accuracy of parameter estimates, since such a model can adequately account for non-ignorable missing values. Whether or not such complex models are actually necessary needs to be investigated in future studies. Simulation studies might aid in evaluating to what extent an inclusion of the missing propensity or relevant covariates can improve parameter estimates. However, our results allow assessing some of the assumptions of other approaches. The fact that the probability for a missing value does not solely depend on the ability in the tested domain refutes the assumption that missing values merely result from lack of knowledge. In low-stakes assessments, missing values should therefore not be treated as wrong. Since the missing mechanism differs for various subgroups and is also different with respect to item format, the assumption of a uniform missing mechanism across all persons and all items does not hold, either.

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

Author note

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

This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 3–5th Grade (Paths through Lower Secondary School - Education Pathways of Students in 5th Grade and Higher), doi:10.5157/NEPS:SC3:2.0.0; This paper uses data from the

National Educational Panel Study (NEPS) Starting Cohort 4–9th Grade (School and Vocational Training - Education Pathways of Students in 9th Grade and Higher), doi:10.5157/NEPS:SC4:1.0.0. This paper uses data from the National Educational Panel Study (NEPS) Starting Cohort 6–Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, the NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research and supported by the Federal States. As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi).

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Holman, R. & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1-17. doi: 10.1111/j.2044-8317.2005.tb00312.x

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

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

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

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

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

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

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

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

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

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

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

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

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

OECD (2012). *PISA 2009 Technical Report.* Pisa: OECD Publishing.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in Item Response

Theory models. *Educational and Psychological Measurement, 74,* 423–452. doi: 10.1177/0013164413504926

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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