PSYCHOMETRIKA—VOL. 84, NO. 3, 892–920
SEPTEMBER 2019
https://doi.org/10.1007/s11336-019-09669-2

USING RESPONSE TIMES TO MODEL NOT-REACHED ITEMS DUE TO TIME LIMITS

STEFFI POHL AND ESTHER ULITZSCH

FREIE UNIVERSITÄT BERLIN

MATTHIAS VON DAVIER

NATIONAL BOARD OF MEDICAL EXAMINERS

Missing values at the end of a test typically are the result of test takers running out of time and can as such be understood by studying test takers' working speed. As testing moves to computer-based assessment, response times become available allowing to simulatenously model speed and ability. Integrating research on response time modeling with research on modeling missing responses, we propose using response times to model missing values due to time limits. We identify similarities between approaches used to account for not-reached items (Rose et al. in ETS Res Rep Ser 2010:i–53, 2010) and the speed-accuracy (SA) model for joint modeling of effective speed and effective ability as proposed by van der Linden (Psychometrika 72(3):287–308, 2007). In a simulation, we show (a) that the SA model can recover parameters in the presence of missing values due to time limits and (b) that the response time model, using item-level timing information rather than a count of not-reached items, results in person parameter estimates that differ from missing data IRT models applied to not-reached items. We propose using the SA model to model the missing data process and to use both, ability and speed, to describe the performance of test takers. We illustrate the application of the model in an empirical analysis.

Key words: item response theory, response times, missing responses, not-reached items, time limit, Bayesian modeling.

1. Missing Values in Cognitive Test Items from Large-Scale Assessments

Large-scale assessments (LSAs) such as the Program for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the National Educational Panel Study (NEPS) aim at accurately measuring competencies such as reading comprehension or mathematical literacy. Competencies in these studies are assessed by tests containing a number of tasks that have to be completed in a certain time. Data collected in LSAs usually show a large proportion of missing responses due to the low stakes nature of the assessment. Missing responses may be due to incomplete block assessment designs (planned missingness), due to item-level nonresponse (omitted responses), or items that were not reached (for example due to time limits). The amount of unplanned missing responses in LSAs is not negligible. In PISA 2006, for example, across all countries and all three domains (mathematics, reading, and science), an average of 10% of the items were omitted and 4% were not reached (OECD, 2009, p. 219–220). Even more important for country rankings, the amount of missing values largely varies across countries (from 1% in the Netherlands to 16% in Kyrgyzstan for omitted items and from 0.3% in Azerbaijan to 13% in Colombia for not-reached items).

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft), Grant No. PO1655/3-1. We thank Wim van der Linden for helpful comments on the manuscript as well as the HPC service of Freie Universität Berlin for support and computing time.

Correspondence should be made to Steffi Pohl, Methods and Evaluation/Quality Assurance, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. Email: steffi.pohl@fu-berlin.de

This relatively large amount of missing responses needs to be dealt with in the psychometric analysis of test data. While not administered items can usually be considered as missing completely at random (MCAR) or missing at random (MAR), omitted and not-reached items (NRIs) are usually nonignorable and may lead to biased estimates of item and person parameters (see, e.g., Lord, 1974; Mislevy & Wu, 1996; Pohl, Gräfe, & Rose, 2014). If not appropriately accounted for, estimates of group statistics can be biased by missing values as well, resulting in, for example, a different country ranking or biased regression coefficients when predicting test performance from explanatory variables (Köhler, Pohl, & Carstensen, 2017, Rose, von Davier, & Xu, 2010). In order to avoid biased item and person parameter estimates, the missing responses need to be appropriately dealt with.

Most of the models for missing values rely on information that is available in paper and pencil tests, such as item responses, item nonresponses, and covariates. However, as testing moves to computer-based assessment (CBA), more information, in particular process and timing data, becomes available. In this article, we bring together research on missing values with research on process data, specifically response times (RTs), and aim to use RT information to model missing values and to describe the performance of test takers. Note that in the following we will focus on missing values due to not reaching the end of the test because of time limits; the proposed approaches are not necessarily suited for other types of missing values.

Note that the target ability that we aim at estimating is *effective* ability (and in addition *effective* speed) as defined by van der Linden (2007). Effective ability is the ability observed at the chosen (effective) speed level. Test takers may and usually do differ in the speed they chose for a given test. Although in substantive and methodological research, this is hardly ever discussed (see van der Linden, 2007, Kuhn & Ranger, 2015, Tijmstra & Bolsinova, 2018 and Pohl & von Davier, 2018 for a few exceptions), this is what is done in almost all competence assessments in large-scale studies. In these assessments, test takers differ in their speed (even if RTs are not recorded) and no adjustment for speed is done.[1] While Tijmstra and Bolsivova (2018) suggest aiming at estimating optimal ability, that is, the ability observed when the test taker uses exactly the time given for answering all test items (i.e., optimal speed), Pohl & von Davier (2018) point out that optimal ability can only be estimated in very specific experimental settings that are hardly feasible in LSAs. They instead suggest estimating effective speed and effective ability as introduced by van der Linden (2007). By doing this, they explicate what has implicitly been modeled in many studies and with many modeling approaches before. It is also in line with the speed-accuracy model of van der Linden (2007). Pohl and von Davier (2018) argue that this approach (a) allows to disentangle the different aspects of performance (a medium performance may be observed for persons with high ability and high speed as well as for persons with lower ability and lower speed), (b) allows to estimate the same target ability for all groups of test takers, and (c) better reflects real-life performance, as persons also need to choose their speed when solving real-world problems outside of testing situations. In this paper we will explicitly follow this approach, i.e., we will focus on effective speed and effective ability making no claims about optimal levels of these.

In the following, we will first introduce research on missing values. Then we will give an overview of research on models for RTs with a focus on the speed-accuracy model of van der Linden (2007). Finally, we will bring these two research lines together and show how RTs may be used for modeling missing data due to time limits. We will also discuss how RTs may be used for describing the performance of test takers in the presence of missing values.

[1] This is different in the study by Goldhammer (2015), who imposed item-level time limits to reduce the heterogeneity in RTs across persons. Note, however, that this only reduces heterogeneity in RTs across persons, but does not get rid of it. Furthermore, item-level time limits may result in guessing and item omission (Kuhn & Ranger, 2015, Pohl & von Davier, 2018).

2. Modeling Missing Values Within IRT

2.1. Classical Approaches

There are different approaches to dealing with missing values (for an overview see, e.g., De Ayala, Plake, & Impara, 2001, or Rose et al., 2010): 1) Missing responses may be ignored and, thus, treated as if they were not administered. This approach assumes that missing responses are MAR, given the observed responses on the items in the test (and other covariates in the background model). This approach is applied to missing values due to NRIs in NAEP (Johnson & Allen, 1992). 2) Missing responses may also be scored as incorrect responses, assuming that the subject did not know the answer. This is a deterministic scoring approach ignoring the fact that any respondent has a positive (even if low) probability to solve any item, given its trait level. Lord (1974) showed that the incorrect scoring method results in biased parameter estimates and proposed 3) to score missing responses as fractionally correct, for example, by scoring them according to the probability of guessing correctly. Fractional correct scoring is used for omitted multiple choice items in NAEP (Johnson & Allen, 1992). 4) In some educational studies (e.g., PISA until 2012, TIMSS, and PIRLS), a two-stage procedure for treating missing responses is used (see, e.g., OCED, 2009). For the estimation of item parameters, missing responses are ignored. The estimated item parameters are then used as fixed parameters for the estimation of person parameters where missing responses are scored as incorrect.[2]

Each of these approaches involves certain assumptions regarding the occurrence of missing responses. These assumptions do not necessarily hold in LSAs (see, e.g., De Ayala et al., 2001; Pohl et al., 2014; Rose et al., 2010). The approaches scoring missing values as incorrect do violate assumptions of IRT models (Lord, 1974; Rose, 2013). Also ignorability (i.e., MAR) of the missing values due to omitted items and NRIs usually does not hold. Different studies (e.g., Glas & Pimentel, 2015; Holman & Glas, 2005; Köhler, Pohl, & Carstensen, 2015; Pohl et al., 2014; Rose et al., 2010) demonstrated that missing responses due to omission and test time limits often depend on the ability of the person and are thus nonignorable.

2.2. Model-Based Approaches for Nonignorable Missing Responses

Recently, model-based approaches for dealing with nonignorable missing data in IRT models have been developed. As these models may account for nonignorable missing data, which is most likely the missing data process present in cognitive test data, we focus on these types of models. In these approaches, the tendency to omit responses is included in the model and accounted for in the estimation of the item and person parameters. The missing response tendency can be either included (1) via a latent missing propensity that is accounted for in a multidimensional IRT model (Glas & Pimentel, 2015; Holman & Glas, 2005; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999) or (2) by defining a manifest missing data indicator that is accounted for in a latent regression or multiple-group IRT model (*manifest missing approach for not-reached items,* mNRI, see, e.g., Rose et al., 2010). Rose (2013) showed that the mNRI model performs well and described the approach as sufficient for NRIs. In the following we will therefore focus on this approach. In the mNRI approach, there is a unidimensional IRT measurement model for the responses U_i to item i. Missing responses due to omissions and NRIs on response variables U_i of item i are treated as missing values in the measurement model of U_i. A missing propensity is computed for each person as the relative number of NRIs $\bar{d}$. This missing propensity is included in the measurement model as an explanatory variable via latent regression, or alternatively, as a grouping variable used in a multiple-group IRT model (see Fig. 1). As such, in the estimation of

[2] This uses item parameters estimated with missing data ignored on data in which missing responses are coded as wrong. Hence, the item parameters do not fit the observed rates of wrong responses. This procedure was abandoned in PISA 2015.

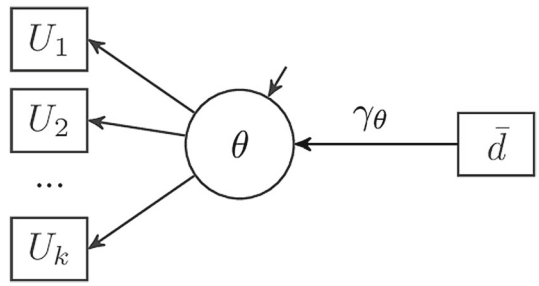

FIGURE 1.
The manifest missing approach for not-reached items.

the item parameters and ability scores of the cognitive measurement model, the relative number of missing responses is controlled for. There is no restriction on the kind of measurement model for the latent ability.

Recently, researchers (Glas, Pimentel, & Lamers, 2015; Moustaki & Knott, 2014; Köhler, et al., 2014; 2015) acknowledged that the missing response process may be even more complex and tried to more accurately model that mechanism by including further covariates explaining the missing data mechanism. Köhler et al. (2015) found that additionally to ability, metacognitive competencies, reading speed, demographic variables (such as immigration background and school type) and interactions of these variables are relevant predictors for the missing propensity. Glas et al. (2015) proposed a model-based approach for dealing with missing values that extends the latent variable approaches for missing values to incorporate further person characteristics.

2.3. The Impact of Response Time

Although the approaches accounting for nonignorable missing responses employ sophisticated modeling and show promising results, they rest on a number of implicit assumptions that may not necessarily hold in practice. The models and the simulation studies carried out in support of the models do not directly consider time on task for solving an item (e.g., Culbertson, 2011; De Ayala et al., 2001; Finch, 2008; Holman & Glas, 2005; Köhler, et al., 2017; Pohl et al., 2014; Rose et al., 2010). However, test takers who do not respond to all items have more time available to solve the items they choose to attempt, compared to test takers who attempt all items. In particular, test takers that do not reach a large number of items have often spent a much longer time on the (few) items they attempted. Studies using real data (e.g., Goldhammer & Kroehne, 2014; Semmes, Davison, & Close, 2011) show that the time available to solve an item does affect the probability of a correct response. This aspect has been, to our knowledge, neglected so far in research on missing response modeling.

3. Response Times Informing the Missing Response Process

Previous approaches for dealing with missing values rely solely on data of responses to test items (and on person characteristics). With the shift from paper and pencil (P&P) assessment to CBA, more information on how the test takers interact with items on a test becomes available. Specifically, CBAs typically collect data on the time test takers use to respond to each item. This information may be valuable for evaluating and modeling missing values in cognitive tests. RTs may provide information about the time allocation strategies of test takers. Moreover, process data do not only provide information on how much time it took to respond to an item, but also on

how much time a person has spent on an item even if the person finally chooses not to respond (nonresponse time). The total time spent on an item (the time point when the item is first displayed to the time point when the respondent moves to the next item), no matter whether a response was produced or not, may help determine whether test takers indeed engaged in solving the item. Very short total time on an item may indicate that the test taker did not make an effort to solve the item, while longer total times make it appear more likely that a person did attempt to solve the item (Weeks, von Davier & Yamamoto, 2016).

Operationally, LSAs such as PISA and TIMSS have not used RTs simply due to the fact that P&P assessments administered to groups of students within schools do not allow for accurate measures. The OECD Programme for International Assessment of Adult Competencies (PIAAC) was one of the first international assessment program that was fully computer-based, in the sense that all test takers except those without sufficient computer experience were tested using laptop computers. The database for this assessment includes timing data and is publicly available. The analysis of RTs and missingness originated in this assessment with a more heuristic rule, based on the observation that missing data appeared to be associated with, on average, much shorter time measures (e.g., Yamamoto, Khorramdel & von Davier, 2013). That means that missing data with a nonresponse time of below a certain threshold was considered to be based on insufficient engagement with a task, while nonresponse that was associated with times usually observed together with incorrect or correct responses was considered as an indicator of a lack of skills. In other words, rapid skipping to the next item was not penalized, and considered a missing response, while respondents who did skip, but after a longer time had passed, would be assigned an incorrect response. This is consistent with findings about rapid guessing in assessments where test takers may feel forced to respond (Wise & DeMars, 2005) based on the high-stakes nature of the assessment. Using data from a large-scale NAEP computer-based field study, Lee and Jia (2014) found that rapid responses have no statistical association with the ability estimated based on responses given when sufficient time has passed. In PIAAC, researchers acknowledge the potential of RT for scoring missing values, however, their approach so far is based on heuristic rules. A sophisticated model that incorporates RT for modeling missing values would strengthen the existing approach.

In PIAAC, the threshold of considering an omitted item as a missing rather than an incorrect response is currently set to 5 s. As research by Weeks et al. (2016) suggests, this rule may establish a lower bound, but item-dependent variability appears to exist. The authors found that RTs vary by item and argue that item-specific thresholds should be chosen. They furthermore showed that a 5-s threshold may be too low and that—depending on the expected probability level of a response—median RT thresholds vary from 7 (expected probability of .50) to 41 s (expected probability of .90).

IRT models have been proposed that utilize RTs as an additional source of information. In these approaches, RTs have mainly been used to model differential speededness of the test. Within this line of research, models have been developed that incorporate RTs in the scaling model to account for differential speed of persons. While these models are quite elaborate, they have neither considered missing values in item responses, nor have they been used to model missing values in item responses. In the following, we will review these models and derive how they may be used for accounting for missing values.

3.1. Response Time Modeling Within IRT

There are different kinds of models that incorporate RTs in the scaling of response data. These either incorporate RTs into the response model, incorporate responses into an RT model, or simultaneously model RTs and responses. An overview of these models is given in Schnipke and Scrams (2002) or Lee and Chen (2011). Note that none of the RT models explicitly deals with

missing values. For our work, we focus on the third class of models, in which RTs and responses are simultaneously modeled. The simultaneous modeling allows depicting the different aspects of testing (ability and speed) separately, but in a combined model. We specifically focus on the speed-accuracy (SA) model proposed by van der Linden (2007). In the following, we describe this approach and discuss its potential utility to model missing values due to time limits in cognitive tests.

3.2. Hierarchical Speed-Accuracy (SA) Model

Van der Linden (2007) notes that, even when accounting for random error, test scores do not automatically reflect the rank order of the test takers' abilities. They do so only when test takers operate at the same speed. Otherwise test takers' scores are "confounded with their decision on speed" (p. 21). The approach of van der Linden clearly distinguishes between ability and speed and recognizes that different persons may choose different speed levels when working on a test. In the model, effective abilities at the chosen (effective) speed of the test takers are estimated.

Van der Linden postulates different characteristics as the basis for his model. First, he proposes that RTs on test items should be treated as realizations of random variables. Second, van der Linden notes that the probability distribution of RTs is different from the distribution of the response variable, but related. Third, RT and speed are not equivalent. Instead, similar to the definition of speed in the natural sciences, speed is defined as the rate of change of some measure with respect to time. As a consequence, RT models with speed as a person parameter should also have an item parameter to quantify varying levels of time intensity. Fourth, speed and ability may be related.

Van der Linden (2007) proposed a hierarchical model that incorporates a structure for simultaneous modeling item responses and RTs. He assumes that response and RT distributions are determined by distinct parameters. At the lower level, the measurement models for item responses and RTs are specified, while at the higher level, the joint distribution of person parameters and the joint distribution of item parameters is specified. For the item responses U_i at the lower level, van der Linden assumes a 3PL model and models the probability of success of an item as

$$p_i\left(\theta_j\right) \equiv c_i + (1 - c_i)\,\Psi\left[a_i\left(\theta_j - b_i\right)\right] \tag{1}$$

with Ψ (.) being the logistic function, θ_j being the ability parameter of test taker j, and a_i, b_i, and c_i being the discrimination, difficulty, and guessing parameters for item i, respectively.

For the RT T_{ij} for each item i and person j the model postulates a lognormal distribution:

$$\ln T_{ij} \equiv \beta_i - \tau_j + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \alpha_i^{-2}\right) \tag{2}$$

where τ_j denotes the speed at which person j operates, β_i denotes the time intensity of the item, and α_i denotes the reciprocal of the standard deviation of the RTs on item i and can be interpreted as a time discrimination parameter. The appropriateness of a lognormal model for RTs has been investigated by van der Linden (2006) and Schnipke and Scrams (1997).[3]

At the higher level, van der Linden postulates parametric distributions for the item as well as for the person parameters of the two lower level models. For the person parameters, he assumes a multivariate normal distribution of the ability and speed variables. For the item parameters, he assumes a multivariate normal distribution for all item parameters in the response model and the RT model (i.e., for a_i, b_i, c_i, α_i, and β_i). The model is depicted in Fig. 2.

[3] See Eq. (04) in their paper.

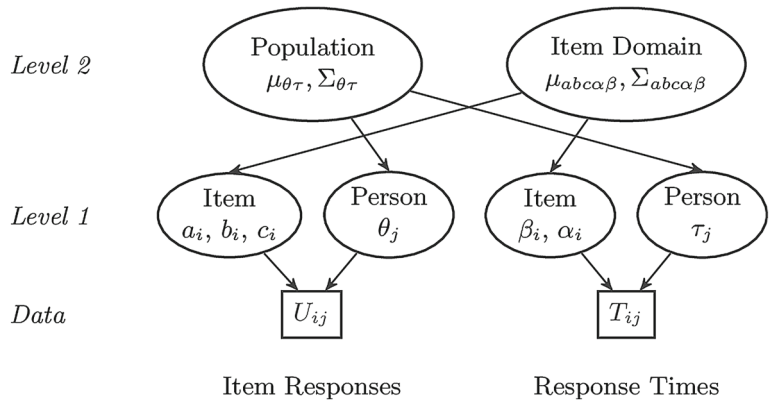

FIGURE 2.
Graphical illustration of the hierarchical model of van der Linden (2007).

Conditionally on ability and speed, the model assumes independence between responses to different items, independence between RTs on different items, and independence between responses and RTs on the same items. Thus, it is assumed that persons operate at constant ability and speed across the test. There may, however, be a dependency of accuracy and speed across persons. This is implemented at the higher level by means of a joint distribution for these random effects allowing for a correlation between speed and ability.

The model of van der Linden has been the basis for further model developments (e.g., Bolsinova, Tijmstra, & Molenaar, 2017; Bolsinova & Tijmstra, 2016; Fox & Marianti, 2016; Molenaar, Oberski, Vermunt, & De Boeck, 2016; Molenaar, Tuerlinckx, & van der Maas, 2015; Meng, Tao, & Chan, 2015; Ranger & Kuhn, 2012; Ranger & Orthner, 2012; van der Linden & Glas, 2010) and substantive research (e.g., van der Linden, 2008; van der Linden & Guo, 2008).

Van der Linden (2007) only mentions in passing the issue of missing values: He notes that in his model "both speed and power aspects [are] captured by the variables T_{ij} (or D_{ij}) and U_{ij}, respectively" (p. 16), where T_{ij} denotes the RT, D_{ij} the indicator missing variable which shows which items the test taker completes, and U_{ij} the response variable. However, he does not further elaborate the role of missing values in his model.

4. Objectives

Missing values occur in cognitive tests and may have a systematic impact on conclusions drawn from cognitive test data. Different approaches for dealing with missing data exist. These rely on information about item responses, missing values, and covariates and try to adequately model the missing data process using this information. While models for RT exist, they have not, yet, been used for modeling missing values. We aim at filling this gap by bringing together research on missing values with research on response times. Specifically, we use the SA model of van der Linden (2007) to account for missing responses due to time limits. We will show the usefulness of the model, discuss implications for the evaluation of the performance of test takers, and give an outlook to further model extensions.

4.1. Using the Speed-Accuracy Model for Modeling Missing Responses due to Time Limits

One may understand the propensity of not reaching items as a measure of working speed. Then, the model of van der Linden is closely related to the mNRI model for NRIs (Rose et

al., 2010; Fig. 1). In both models, ability and a measure of speed are considered. Both models include two variables for speed and accuracy simultaneously, but separately, and the relationship of ability and speed (or the tendency to [not] reach items) is estimated. Both models make the assumption that ability and speed/missing propensity are stationary within the test. They are also similar in the (implicit or explicit) assumption of independence of responses and RTs or missing indicators given ability and speed. Note that in contrast to Rose et al. (2010), van der Linden (2007) also estimates a measurement model for speed as well as the joint distribution of the item parameters of the response and the RT model.

While in the model by Rose et al. (2010) speed is indicated by the number of NRIs, indicators for speed in van der Linden's model are the RTs per item. The number of NRIs can be seen as a rough proxy for RT at the level of the whole test. Thus, the information contained in the number of NRIs is also contained in the RTs. The RTs provide even more detailed information, that is, information on how much time a person has spent on *each* item. Thus, the model by van der Linden may be able to account for nonignorable missing values due to time limits.

4.2. Investigating the Performance of the Model and Comparing it to Previous Models

In this paper, we investigate whether the SA model of van der Linden may indeed be sufficient to account for missing values due to not reaching the end of the test because of time limits. We investigate the performance of the SA model in comparison to the model by Rose et al. for accounting for missing values. If the SA model proves suitable, this approach may not only help to model missing values due to time limits, but may also provide information about the missing response process. Furthermore, it may help bringing together the research traditions of modeling missing values and those of modeling response times.

5. Method

We conducted a simulation study in which item responses and missing values were generated following the SA model. For data analyses, we applied the SA model as well as the mNRI model by Rose et al. (2010). Note that as the SA model is used as the data-generating model, it is the true model with respect to comparisons of models made in subsequent analyses. We decided to use the SA model as the data-generating model for two reasons: First, the SA model is the more informative model from which data of both approaches may be generated, i.e., it is not possible to generate RTs from the number of NRIs alone without further assumptions. Second, and more important, missing responses due to time limits of the test are thought to be determined by the total of the times taken for each item. That is why the SA model will allow incorporating a theoretically sound missing data mechanism. As a consequence of this simulation design the SA model will fit the generated data better than the mNRI model. However, we do not aim at such a comparison, but rather at (a) evaluating whether the SA model is able to correctly estimate ability parameters in the presence of nonignorable missing data and (b) investigating whether the SA and the mNRI model result in similar parameter estimates, in particular, to what extent ability estimates agree between these approaches.

5.1. Data Generation

For data generation, we chose parameters that represent typical low stakes LSAs. Employing the SA model as the data-generating model, we generated data for $N = 1000$ persons responding to $K = 30$ items. For setting the parameters of the SA model, we relied on empirical results from the application of the SA model to emprical data, while the applications were not specifically considering missing values (e.g., Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2007;

van der Linden, Breithaupt, Chuah, & Zhang, 2007; van der Linden & Guo, 2008; van der Linden, Scrams, & Schipke, 1999).

For the person parameters, θ and τ, we chose the following settings: $(\theta, \tau) \sim MVN(\mu_P, \Sigma_P)$ with $\mu_P = (0, -3.5)$ and $\Sigma_P = \begin{pmatrix} 1 & \text{cov}(\theta, \tau) \\ \text{cov}(\theta, \tau) & 0.25 \end{pmatrix}$. This corresponds to findings from empirical data, for example, in van der Linden, Scrams and Schnipke (1999). The correlation between the person parameters cor (θ, τ) varies a lot in empirical data, ranging from cor $(\theta, \tau) = 0.30$ (van der Linden et al., 2007; van der Linden, 2007) to cor $(\theta, \tau) = .04$ (van der Linden et al., 1999) to negative values down to cor $(\theta, \tau) = -.76$ (Klein Entink et al., 2009). We reflected this range of results by choosing different levels of correlations in our simulation, that is, cor $(\theta, \tau) = (-.50, .0, .50)$.

As the measurement model for item responses, we chose the Rasch model, as it is used, for example in PISA until 2012, or in NEPS. Hence, only item difficulties b_j need to be estimated. For the measurement model of RTs, we fixed the discrimination $\alpha_j = \alpha = 1.875$ to be the same across all items. The value 1.875 was chosen in accordance with empirical results of van der Linden (2006). For the remaining item parameters, b and β, we assumed a multivariate normal distribution: $(b, \beta) \sim MVN(\mu_I, \Sigma_I)$ with $\mu_I = (0, 0)$ and $\Sigma_I = \begin{pmatrix} 1 & \text{cov}(b, \beta) \\ \text{cov}(b, \beta) & 0.14 \end{pmatrix}$. In empirical data, correlations cor (b, β) between difficulty and time intensity of items have been found to vary between cor $(b, \beta) = .30$ (van der Linden, 2007) and cor $(b, \beta) = .51$ (Klein Entink et al., 2009) and can even be as high as cor $(b, \beta) = .65$ (van der Linden et al., 1999). We reflected this range of correlations and added a zero correlation as a reference in our simulation design, resulting in simulated correlations of cor $(b, \beta) = (0, .60)$.

We combined each of the two item parameter correlations with each of the three person parameter correlations, resulting in six simulation conditions. For each condition, we generated 100 replicate datasets. The person and item parameters were fixed and used in all 100 replications.

We used the formulas in Eqs. 1 and 2 to generate item responses and RTs for each replication in each condition based on the generated item and person parameters. This resulted in $9^*100 = 900$ complete data sets without any missing values. The median RT of the items ranged from 14.17 to 71.58 s. Within this data generation, no time limits were assumed.

We then considered a test setting in which the time limit for the test is 30 min. This corresponds to the usual time limit of test forms in LSAs (e.g., in NEPS, test forms ranging from about 24 to 36 items are presented with a 30 min time limit, see e.g., Duchardt & Gerdes, 2012; Pohl, Haberkorn, Hardt, & Wiegand, 2012; Senkbeil & Ihme, 2012). We induced missing values based on the cumulative RT of the items. The items were assumed to be in the same order for every person (e.g., as in NEPS, Pohl & Carstensen, 2012) and the RTs were cumulated across the position of the items in the test. All items with a cumulated RT exceeding the time limit were assumed to be not reached and hence responses to these items were coded as missing. This resulted in incomplete data sets with 4.73 to 5.71% of missing values.

We subsequently assessed the effects of sample size, number of items, as well as rate of NRIs. To do so, we chose the condition with cor $(\theta, \tau) = .50$, cor $(b, \beta) = .60$, as this was one of the conditions with the most severe threat to parameter estimation. We varied the number of examinees (adding a condition of $N = 500$), the number of items (adding a condition of $K = 10$), and the rate of NRIs (adding a condition of 15%). We controlled the amount of NRIs by setting stricter time limits (1200 s for a missing rate of 15% under conditions with $K = 30$; 600 and 400 s for missing rates of 5% and 15%, respectively, under conditions with $K = 10$). This resulted in seven additional conditions (two sample size conditions times two item number conditions times two rate of NRIs conditions minus the baseline condition) evaluated in additional analyses.

5.2. Data Analysis

We analyzed the generated data using the SA model as well as the mNRI model. First, in order to evaluate how the SA model deals with missing values in general we applied the SA model (a) to the complete data (SAcomp) as well as (b) to the incomplete data (SAinc) and compared the results. Second, in order to evaluate the difference between the SA model and the mNRI model for dealing with missing values, we applied both to the incomplete data. In order to get comparable results, we estimated all models using Bayesian estimation with Gibbs sampling in JAGS (Plummer, 2003), making use of the 'rjags' package (Plummer, 2016) in R version 3.3.2 (R Development Core Team, 2016). Missing values in JAGS are imputed based on the specified model. We used non-informative priors, keeping the priors for the same parameters constant in all models. The settings for priors and syntax for the analyses can be found in Appendix A.

For both, the SA model and the mNRI model, we used three chains and no thinning. For the SA model, we chose a total of 45,000 iterations per chain, with a burn-in of 5000, yielding a total of 120,000 iterations for posterior analyses. For the mNRI model, we ran 15,000 iterations per chain using a burn-in period of 5000; altogether 30,000 iterations were saved.

We evaluated convergence of the model via trace plots and the Gelman–Rubin Potential Scale Reduction Factor (PSRF, Gelman & Rubin, 1992). We checked autocorrelation by assessing plots of the autocorrelation function along with the effective sample size (ESS as described in, e.g., Drummond, Nicholls, Rodrigo, & Solomon, 2002). For evaluating the performances of the models in retrieving accurate parameter estimates, we examined the estimates of person ability and speed, item parameters, as well as the correlation between ability and speed and between item difficulty and time intensity.

6. Results

In the following, we will first present the results for conditions with $N = 1000$, $K = 30$, and a rate of NRIs of 5%. We will then show the results of the effects of sample size, number of items, and rate of NRIs.

6.1. Convergence and Efficiency

Across all conditions and models, no convergence problems were encountered. All trace plots indicate good mixing of the chains and convergence. For both models, PSRF values remained far below 1.05 for all parameters and thus were, in line with Gelman and Shirley (2011), considered acceptable.

Autocorrelation in the MCMC chains varied largely across parameters when data were analyzed with the SA model. ESS ranged from 473 and 526 (for a time intensity parameter) to 26,691 and 26,694 (for an item parameter variance estimate) when the SA model was applied to complete and incomplete data sets, respectively. This indicates that parameter space had been sufficiently explored to assess posterior means and standard deviations (Kruschke, 2014, p. 184). For the mNRI model, ESS ranged from 5587 (for a difficulty parameter) to 10,653 (for an ability estimate).

6.2. Performance of the Speed-Accuracy Model for Complete Data

As expected, the SA model for complete data yielded unbiased group-level parameter estimates across all conditions with bias for all parameter types remaining below 10%. Figure 3 shows the difference of the individual ability estimates averaged across all replications in the SA model for complete data (SAcomp) from the true parameters for the nine conditions. There is a

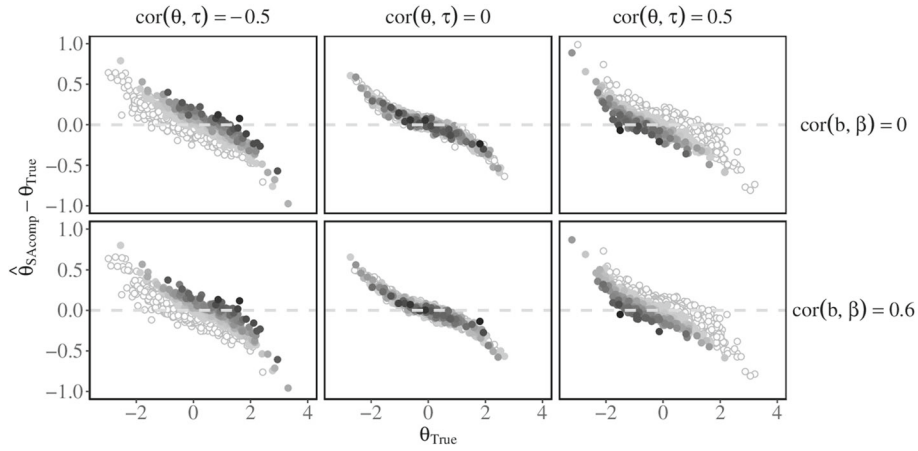

FIGURE 3.
Difference in ability estimates using the SA model for complete data compared to the true ability values as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

noticeable shrinkage effect (see also Fox, 2010) in all conditions. Although no missing values were induced for these analyses, in the graphs we marked the average number of NRIs for each simulee to be induced later. The estimates do not systematically differ across conditions with different item parameter correlations. There are, however, differences across conditions with different person parameter correlations. This is due to the adjustment that is made by incorporating speed in the model. For a person parameter correlation of zero, there is no systematic difference in person parameter estimation of persons with different numbers of missing values. For a correlation unequal to zero, the relationship of the difference in person parameter estimation with the speed variable becomes evident. A negative correlation means that more proficient test takers work slower and therefore have a tendency to produce more missing values. Also, for the same ability level, slower test takers are those that have more missing values (see Figs. 3 and 4) and by the negative correlation between ability and speed this results in a higher ability estimate. Note that in the SA model for complete data, there are no missing values; Figs. 3 and 4 only show the missing values that w*ill be* induced. As can be seen in Fig. 4, there is no systematic bias in ability estimation for different values of true speed. We found similar effects for the estimation of speed. There is a shrinkage effect in the estimation of speed due to unreliability and there is an effect of the correlation between ability and speed (Figs. 14 and 15 in Appendix B). Summarizing the results, it can be concluded that, in the case of complete data, the SA model was able to adequately recover the true parameters. That is, the complete data model may serve as a reference for comparison for subsequent analyses.

6.3. Performance of the Speed-Accuracy Model to Deal with Not-Reached Items

There is no systematic bias in group-level parameter estimates of the SA model for incomplete data, bias for all parameter types remained below 10%. In the following, we compare the results of the SA model applied to incomplete and complete data, respectively. By doing this, we control the shrinkage effect due to using only 30 items, as this number is the same in both cases. Differences between the results of the two analyses can, thus, be attributed to the existence of missing values. Figure 5 shows the difference in ability estimates averaged across all replications between the SA model for incomplete (SAinc) and the SA model for complete data (SAcomp) as a function of true ability and the number of missing values in all conditions. The results show that for simulees

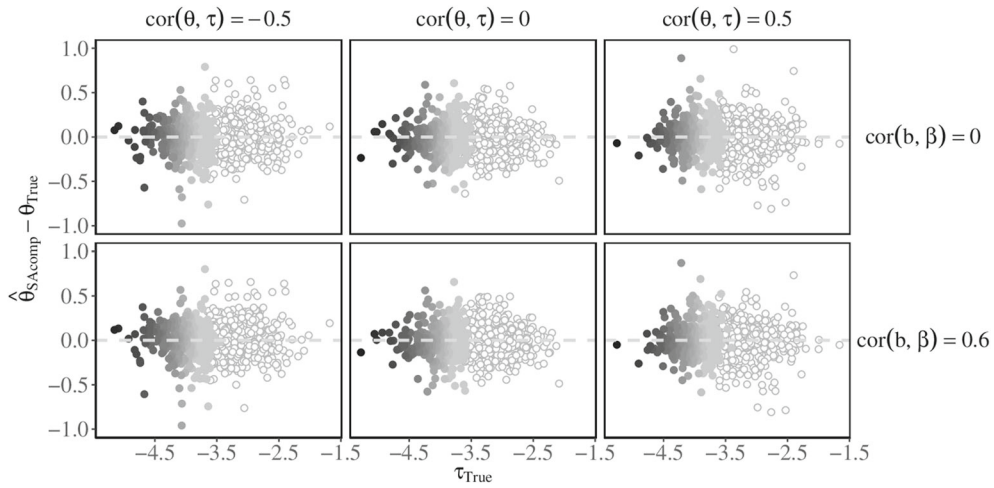

FIGURE 4.
Difference in ability estimates using the SA model for complete data compared to the true ability values as a function of true speed. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

with no missing values, there is no difference in parameter estimates. There is a difference in parameter estimates for simulees with missing values; with greater differences being observed for simulees with a higher number of NRIs. This difference may be explained by a shrinkage effect that is due to the fact that for these respondents, information from fewer than 30 items is available. This is reflected in the posterior standard deviation for simulees with a high number of NRIs: for persons with an average number of NRIs greater than 15, posterior standard deviations ranged from 0.36 to 0.89—as compared to a range from 0.36 to 0.43 for the same person parameter estimates, however, estimated using complete data. The greater uncertainty of these estimates is associated with an increased shrinkage effect.

Person parameters are thus estimated closer to the overall ability mean. This effect is even more obvious when plotting the difference in ability estimates as a function of speed (see Fig. 6). Since simulees with a lower speed produce more missing values, fewer item responses are available, resulting in greater standard errors of ability estimates and a larger shrinkage effect. We found similar results for the estimation of speed (see Figs. 16 and 17 in Appendix B)

6.4. Illustrating the Shrinkage Effect due to Missing Values

In order to underpin the conclusion that the difference in parameter estimates between the SAcomp and SAinc model go back to shrinkage effects due to missing values, we re-ran the analyses on the performance of the SA model for incomplete data with missing values induced completely at random (SA MCAR). To do so we used the complete data sets of 1000 examinees responding to 30 items and introduced missing values being MCAR. The number of missing values for each person was drawn from a discrete uniform distribution ranging from 0 to 25. The number of missing values for each person was fixed across replications; however, the specific items bearing missing values were chosen randomly for each replication. By doing so, we ensured that the average number of missing values for each person across replications displayed a range comparable to the range of the average number of NRIs in the main simulation. Figure 7 depicts the difference in ability estimates between using the SA model on a data set with missing values being MCAR and the SA model on the complete data set as a function of true ability and the

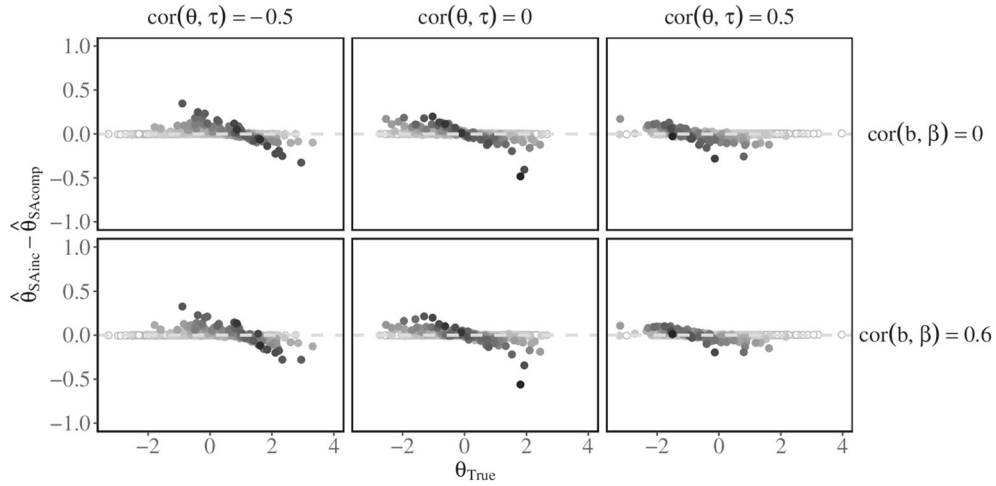

FIGURE 5.
Difference in ability estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true ability. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

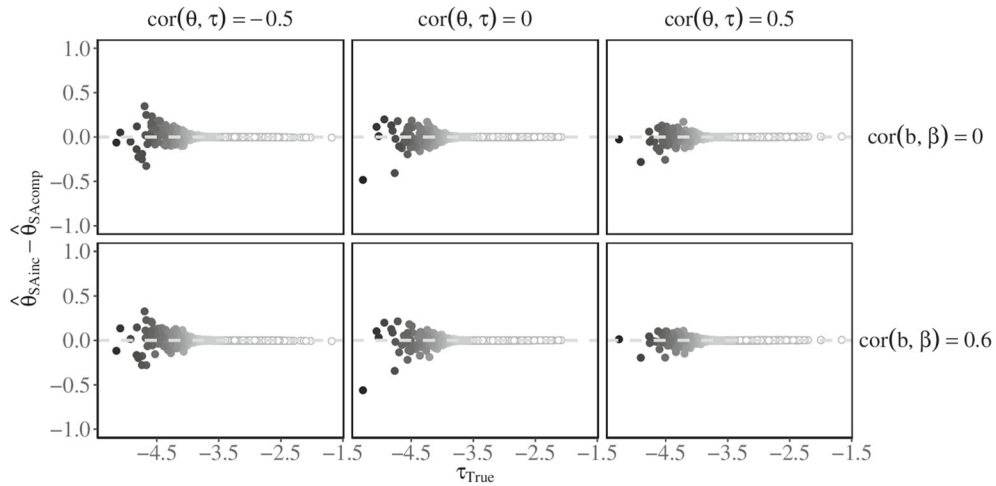

FIGURE 6.
Difference in ability estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true speed. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

number of missing values for the condition with cor $(\theta, \tau) = .50$ and cor $(b, \beta) = .60$. As can be seen, the resulting pattern resembles the pattern of differences between the SA incomplete and the SA complete model, where the difference increases as a function of the number of missing values. Since missing values for the SA MCAR model were induced completely at random, systematic bias in person parameter estimates can be ruled out as an explanation for these differences; instead, the differences can be attributed to the shrinkage effect as a result of the reduced amount of information for individuals with fewer observed responses and a high number of missing

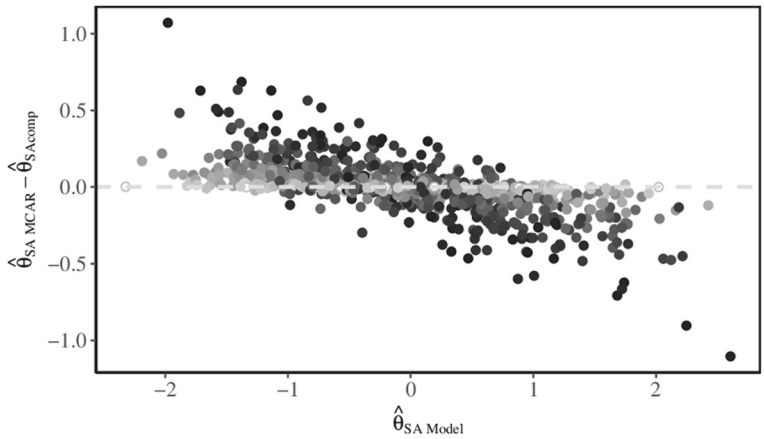

FIGURE 7.
Difference in ability estimates between the SA model for incomplete data with missing values induced completely at random (SA MCAR) and the SA model for complete data (SAcomp) as a function of true ability for the condition with cor $(\theta, \tau) = .50$ and cor $(b, \beta) = .60$. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

responses. From these results, we conclude that the SA model can account for missing values due to time limits. In the next step, we compare the SA model and the mNRI model for dealing with missing data.

6.5. Performance of the Manifest Missing Response Model for Incomplete Data

The simulation results show that bias in hyperparameters is comparable to the SA model. There are, however, differences in person parameter estimation. So far, in the analyses we have found (a) a shrinkage effect due to using only 30 items and (b) a shrinkage effect due to missing values. These are to be expected for this model as well, as it uses the same data as the SA model for incomplete data. In order to separate the shrinkage effects from differences between the models, we compared the mNRI model to the SA model for incomplete data (which was shown to appropriately recover the parameters of the model). Figure 8 and 9 shows the difference in ability estimates averaged across all replications between the mNRI model and the SA model for incomplete data. There is no impact of item parameter correlation. There is, however, one of person parameter correlation. For cor $(\theta, \tau) = 0$, there is no difference between the two models. For a correlation of cor $(\theta, \tau) \neq 0$, there are noticeable differences between the two models, which depend on the true ability and on the number of missing values.

There are two mechanisms at work here: First, as all persons with sufficient speed reach all items (i.e., have no missing values), there is a truncation of the distribution of speed: The manifest missing variable does not distinguish between persons who work with sufficient speed; all of these persons have zero NRIs. As such, for cor $(\theta, \tau) = -.50$, the ability of persons with a high speed level is overestimated, while it is underestimated for cor $(\theta, \tau) = .50$. This is due to the adjustment made to ability due to working speed. As in the mNRI model persons with high speed have the same number of not reached items (i.e. zero), these persons are—in contrast to the data-generating model—treated the same. In the data-generating model, for cor $(\theta, \tau) = .50$ persons with high speed are those with high ability; in the mNRI model, this is not accounted for on the upper speed level.

Second, the mNRI model underestimates the association between ability and speed. The estimated correlation of the number of NRIs (as a proxy for speed) and ability is on average .32,

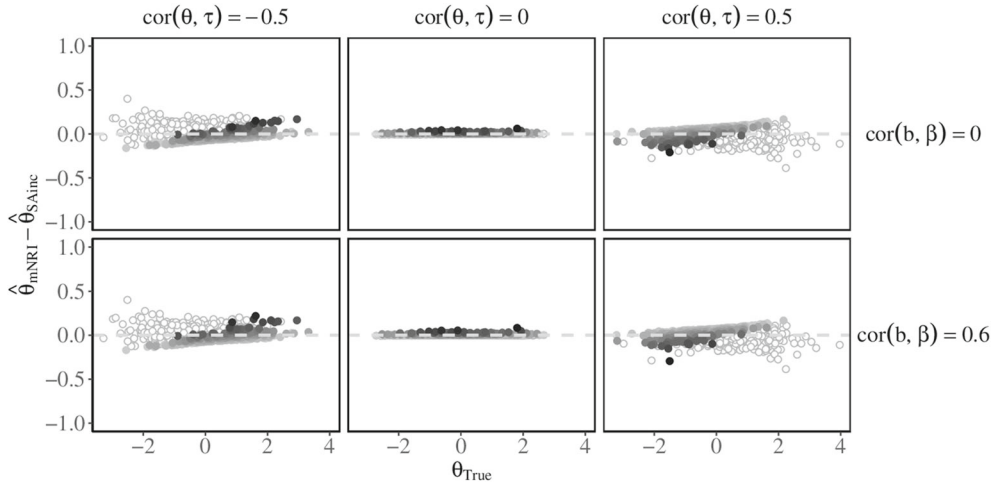

FIGURE 8.
Difference in ability estimates from the mNRI model and the model for incomplete data (SAinc) as a function of ability. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

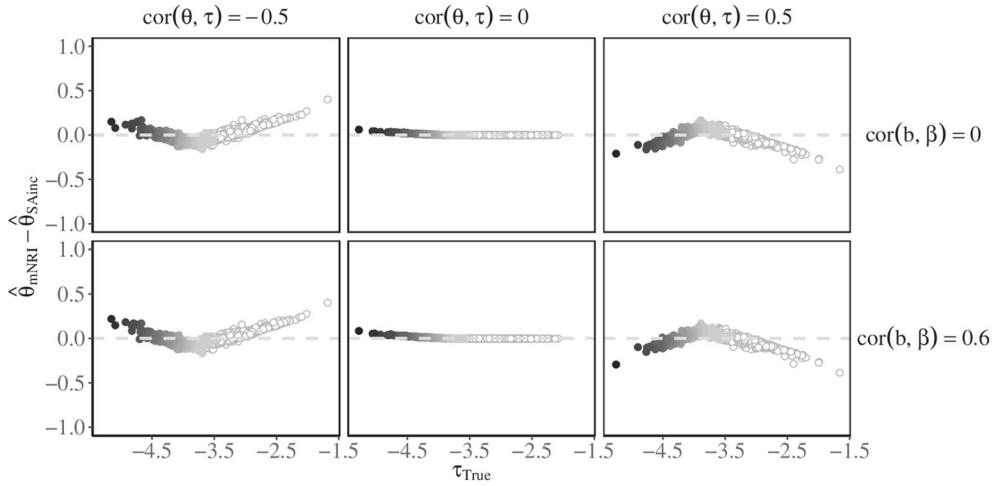

FIGURE 9.
Difference in ability estimates between the mNRI model and the SA model for incomplete data (SAinc) as a function of true speed. White circles represent simulees without missing values and filled circles simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

.02 and −.33 in the condition with cor $(\theta, \tau) = -.50$, 0, and .50, respectively. This may be due to the fact that (a) a manifest instead of a latent variable is used (see Pohl et al., 2014, for a similar result using manifest and latent missing propensity) and/or (b) there is truncation and as such a variance reduction, and/or (c) the number of NRIs is a nonlinear transformation of speed (see Fig. 10) which may violate the linearity assumption of the relationship of ability and speed.

As a consequence, the adjustment of ability estimates based on speed is different than in the SA model for incomplete data. In Figs. 8 and 9, we do see that ability estimates of respondents with low speed (i.e., having many missing values) is overestimated (underestimated) in the condition of cor $(\theta, \tau) = -.50$ (cor $(\theta, \tau) = .50$).

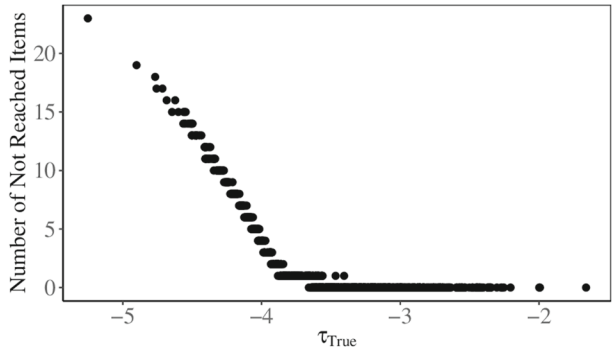

FIGURE 10.
Number of not-reached items per person across as a function of true speed for the condition with $N = 1000$, $K = 30$, cor $(\theta, \tau) = .50$, cor $(b, \beta) = .60$, and a rate of not-reached items of 5%.

6.6. *Additional Analyses*

We conducted additional analyses to study the effects of sample size, number of items, as well as rate of NRIs on parameter estimation. Both models, the SA model for incomplete data and the mNRI model, converged across all conditions and replications. Again, no systematic bias was found in group-level parameter estimates of the SA model. Only for conditions with a small number of items ($K = 10$), item parameter variances were estimated substantially higher than the respective true values (1.11 as compared to 1.00 and 0.25 as compared to 0.14 for var (b) and var (β), respectively). Comparable effects on item parameter variance estimates occurring for smaller number of items have been reported by Fox and Marianti (2016). Appendix C shows the result of person parameter estimation using the SAcomp model as compared to the true ability parameters (Fig. 18) and using the SAinc model as compared to the SAcomp model (Fig. 19). There was no effect of sample size. As was to be expected, an increase in the amount of missing values resulted in an increased shrinkage effect as less information was available for persons with more missing values. There was also an effect of the number of items: There were larger shrinkage effects for conditions with $K = 10$, since under these conditions less information on the examinees' ability is available.

Figure 11 shows the difference in ability estimates when using the mNRI approach as compared to the SA model for incomplete data. Again, there is no effect of sample size. There is, however, an effect of the number of items: with more items, there is more information available for estimating person ability. As a result, the impact of speed becomes smaller relative to the impact of item responses. There is also an effect of the amount of missing values, with more missing values resulting in smaller differences between the two approaches. This is an effect of an increase in discrimination regarding differences in speed and a reduction in the truncation effect in the mNRI model. With an increasing rate of NRIs, the number of NRIs displays more variation and, as such, differences in speed are reflected better by the number of NRIs. As a consequence, the relationship between speed and ability can better be captured in the mNRI model by the relationship of number of the NRIs and ability. Whereas in conditions with a rate of 5% NRIs the average correlation between ability and the number of NRIs was as low as $-.35$ as compared to the generated correlation between ability and speed of cor $(\theta, \tau) = .50$), in conditions with a rate of 15% NRIs an average correlation of $-.42$ was estimated.[4]

[4] These results refer to a condition with $N = 30$ items and $N = 1000$ persons.

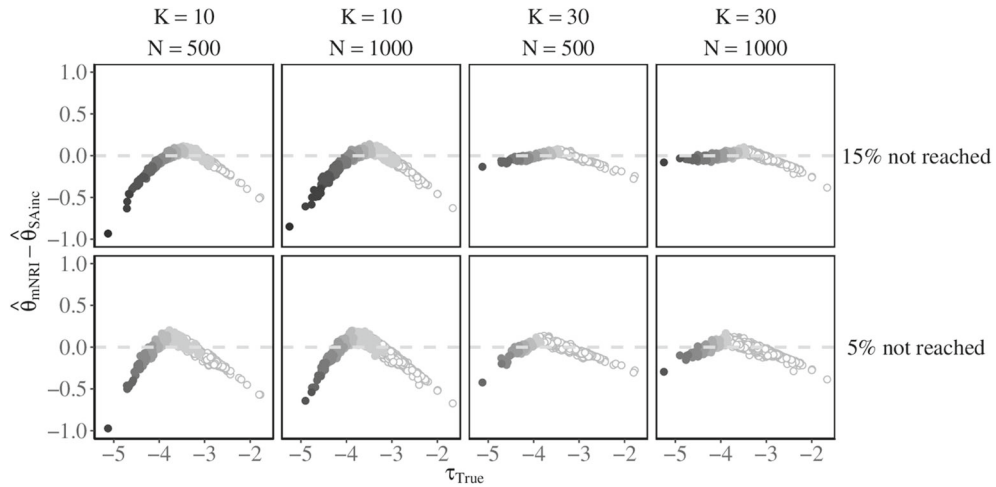

FIGURE 11.
Difference in ability estimates between the mNRI model and the SA model for incomplete data (SAinc) as a function of true speed. White circles represent simulees without missing values and filled circles represent simulees with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

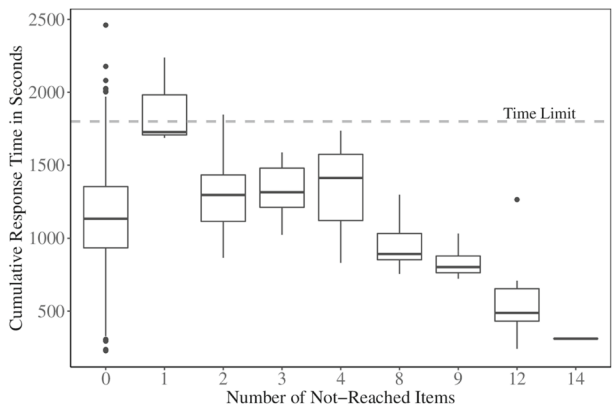

FIGURE 12.
Cumulative response time distribution for different numbers of not-reached items.

7. Empirical Data Analysis

In order to evaluate the applicability of the approach, we analyzed data from the Canadian sample of PISA 2015 (OECD, 2017). We applied (a) the mNRI model and (b) the SA model to science cluster number 7 administered in the second position of the computer-based assessment. In total, analyses were based on $N = 840$ examinees responding to $K = 17$ items within a time limit of 1800 s. Six percent of the test takers did not reach all items. In total, the science cluster under consideration displayed a rate of NRIs of 2%. For reasons of simplicity, partial credit items were dichotomized and examinees with missing data other than NRIs were removed from the analyses. We analyzed the data employing the same MCMC setup as in the simulation study described above, saving 6000 and 8000 iterations as a sample of the posterior distribution for the

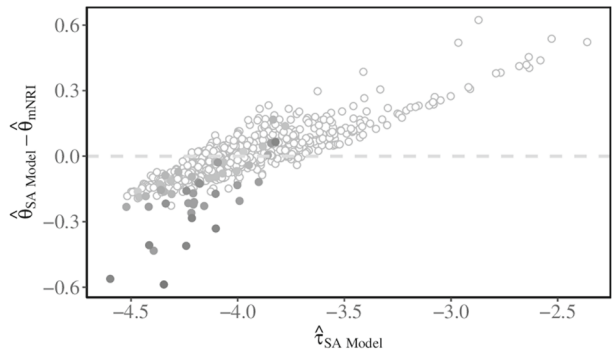

FIGURE 13.
Difference in estimated ability scores between using the mNRI model and the SA model in the empirical application. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

mNRI model and the speed-accuracy model, respectively. Convergence was assessed on the basis of trace plots as well as PSRF values of all parameters. Judging by these criteria, both models converged.

Figure 12 shows that not all test takers stopped working on the test when the test time limit of 1800 s was reached. This indicates that the time limit may not have been enforced rigorously. Most importantly, we see that only for persons with one or two NRIs, the time limit was reached. Almost all test takers with more than one NRI did not use the time they had. Thus, the test time limit seems to be not the only reason for NRIs in the test, but missing values at the end of the test may also occur due to quitting. Note that speed as estimated in the SA model and the number of NRIs correlate −.16. This reflects the results from the descriptive statistics, indicating that NRIs do not only occur due to low speed and reaching the time limit.

Figure 13 shows the difference in ability estimate between using the mNRI model and the SA model. For persons without missing values, we see a similar pattern as in the simulation study. As NRIs are a result of time limits for only very few test takers, for persons with missing values, this pattern deviates from the pattern in the simulation study (see, e.g., Fig. 9). This can also be seen in the parameter estimates of the two models. Using the SA model, we estimated a correlation of .40 between ability and speed. The correlation of ability and number of NRIs in the mNRI model was estimated to be −.01. Summarizing the results, the SA model can account for NRIs that occur due to reaching the time limits. However, NRIs did not only occur due to time limits in this data set.

8. Discussion

In this study, we integrated research on RT modeling with research on modeling missing responses. We proposed using the SA model to model and account for missing values due to time limits in the test. Our study showed that there are similarities between the mNRI model for NRIs (Glas & Pimentel, 2015; Rose et al., 2010) and the SA model of van der Linden (2007). In particular, we identified the potential of the SA model to account for missing values due to time limits on tests. In a simulation study that models this scenario, we showed (a) that the SA model can recover parameters in the case of missing values due to time limits and (b) that the SA model results in different person parameter estimates than the mNRI model. If missing values due to not reaching the end of the test occur because test takers work at different speed levels, then the SA

model can describe the missing data process. In addition, the SA model incorporates differences in working speed also for those test takers who do reach the end of the test (i.e., who have no missing values).

Note that we explicitly aimed at estimating *effective* ability and *effective* speed. Even for test takers without missing values, using the SA model or even just a unidimensional model for responses, effective ability is estimated. The approach adopted here for NRIs due to time limits also estimates effective ability for test takers with missing values, resulting in the same target ability for all groups of test takers, those with and those without missing values.

Of course, we cannot determine with certainty whether the assumed missing data mechanism is the one at work leading to different numbers of NRIs for respondents with different working speeds. From a theoretical point of view, the mechanism seems quite plausible and in the real data analyses we found evidence in supporting of this mechanism for some test takers. While the mNRI model does not describe the mechanism of how missing values occur and includes the missing propensity for adjustment purposes only, the SA model describes a mechanism that explains missing values in terms of time spent on previous items. Empirical studies using response times do hint at the plausibility of such mechanisms (e.g., Goldhammer & Kroehne, 2014).

Note that in our simulation study, we only considered NRIs that occur due to time limits. We did not consider NRIs that occur because the test taker quits responding before reaching the last item. In the empirical analysis, we found evidence that this is another plausible mechanism in practice. While NRIs due to time limits can be accounted for by the SA model, early quitting behavior needs to be accounted for differently. RTs and other log data provide comprehensive information that may help to distinguish between different nonresponse mechanisms.

Moreover, the SA model proposed by van der Linden assumes stationarity of speed and is most appropriately applied to data stemming from tests with a generous time limit. This, however, is not necessarily the case in LSAs that administer tests to groups of students: Testing situations in which test takers encounter (tight) time restrictions and are either running out of time, or perceive it to be so, might lead some test takers to speed up towards the end of the test in order to finish the test within the allocated time. When working speed is used to account for missing values due to NRIs, the very fact that some test takers were not able to reach the end of the test is an indicator that testing time has been not sufficient for all participants. As a consequence, some test takers might have adjusted their working speed in order to reach the end of the test (e.g., Yamamoto & Everson, 1997). Under these conditions, the assumption of stationarity of speed is not plausible and it appears necessary to allow within-person variation of speed (Fox & Marianti, 2016; Goegebeur, De Boeck, Wollack, & Cohen, 2008).

Just by the position of a missing response in the test, within the test (omitted) or at the end of the test (not reached), one cannot infer whether the item had really been attempted or not. It may well be that some omitted items within the test have not been attempted at all (resulting in low nonresponse time) or that NRIs at the end of the test have been attempted (resulting in higher nonresponse time for these items). So far, models for missing values have relied on the position of the missing items within the test for drawing inferences on whether the item had been attempted or not. In some LSAs (e.g., NAEP, Allen et al., 2001), the treatment of missing values is even based on this distinction (omitted items are scored partially correct and NRIs are treated as if they were not administered). By using RTs, one could potentially infer better whether items with missing responses have been attempted or not, compared to making this determination just based on the position of the missing response within the test. How not attempted items within a test and different kinds of missing values can be evaluated with the help of RTs is one important and promising future research task. While Weeks et al. (2016) set out to explore how this can be achieved, their study remains mainly descriptive. A more model-based approach is needed that describes mathematically the interdependence between the time spent and the time remaining on the one hand and response vs. omission propensity on the other hand. The present study shows

that the SA model proposed by van der Linden, together with some assumptions about how time on tasks and time limits relate to not-reached responses, can be used to model data by utilizing more information than the missingness indicators alone.

8.1. Implications for the Practice of Dealing with Not-Reached Items due to Time Limits

In low stakes LSAs—as they are currently implemented—persons differ in their working speed. As a consequence, test takers are on a different position with regard to their speed-accuracy trade-off (van der Linden, 2007). As such, we do estimate the *effective* ability and *effective* speed, which differs between test takers. This is true no matter whether we actually assess RT or not, or whether persons reach the end of the test or not. Unless in very specific experimental settings (Goldhammer 2015), which are however not feasible in LSAs, it is not possible to correct for chosen speed and to estimate optimal ability (i.e., the ability observed when speed is chosen so that the exact given testing time is used; not more or less). Thus, in line with van der Linden (2007), we argue for the estimation of effective ability, as this quantity can be estimated in the majority of testing situations. With the use of RTs and by modeling the association between speed and ability, we can describe these different aspects of performance.

We suggest describing the performance of groups of test takers (for example grouped by language, country, or school type) by all aspects of performance: ability *and* speed (and/or missing propensity in case of other reasons for missing values) and use all of these for evaluating the performance (see also Pohl & von Davier, 2018). This allows to develop a richer description of differences in performance and to disentangle the different constructs involved. For example, Cosgrove and Cartwright (2014) investigated the decrease in the PISA trend results of Ireland from 2003 to 2009 and found that students showed much larger amounts of missing values in later PISA assessments. They concluded that the decrease in PISA score may be a result of lack of motivation that led to more omitted responses. If the different aspects, that is, effective ability and speed (and/or possibly missing response propensity) would have been estimated and presented separately, these changes over time would have been more evident, and the apparent performance differences could have been understood in the light of other changes (see Sachse, Mahler, & Pohl, 2019 for an investigation thereof).

When comparing, for example, the performance of different countries in a cognitive domain, one may want to compare these on both effective ability *and* effective speed. For country rankings, policy makers are interested in the comparison of only one score for each cognitive domain. If a single score is of interest, we suggest using a composite score based on the estimated aspects of performance. Substantive researchers may then decide how to combine ability and speed estimates by developing a composite score that reflects the dimension they want to focus on most. One advantage of such an approach would be that this composition of a total score would be the same for all test takers. Furthermore, the approach can also deal with varying time restrictions, as has been present in the PISA data. As the measure of speed does not dependent on the total time used, more or less rigorous enforcement of time limits may be accounted for.

This is different in the approach of scoring missing values as incorrect. Scoring missing values due to not reaching the end of the test as incorrect is also a constructed measure incorporating accuracy of responses and speed into a single score. However, (a) the different aspects of performance cannot be disentangled. As such, subgroups with the same estimated average score may differ in effective ability and effective speed. The score may be a result of high effective ability or high effective speed. (b) Speed is only corrected for in the scoring for persons that do reach the time limit, but not for test takers that complete the test within the limit. However, even test takers that are within the time limit differ in their speed. Thus, different target abilities would be estimated for both groups (see also Pohl & von Davier, 2018). This is not the case in the SA model. If speed should be part of the construct to be measured, it should be incorporated in the

same way for all test takers. Additionally, c) scoring NRIs as incorrect results in violations of model assumptions (local stochastic independence, measurement invariance) and was recognized to introduce bias more than 40 years ago (Lord, 1974). Last but not least, d) differences in test time limits are not accounted for by incorrect scoring. Thus, we think that it is valuable and incorporates the different advantages of the previous approaches to first disentangle the different aspects of performance using the SA model and building composite scores in an additional step.

Appendix A

Prior settings	
Speed-accuracy model	$\Sigma_P \sim IW_{2+1}(I_2)$
	$\Sigma_I \sim IW_{2+1}(I_2)$
	$\mu_b \sim N\left(0, 1000^2\right)$
	$\mu_\beta \sim N\left(1, 1000^2\right)$
	$\alpha^2 \sim \Gamma(0.01, 0.001)$
Manifest missing response model	$\sigma_\theta^2 \sim IG(0.01, 0.001)$
	$\gamma \sim N\left(0, 1000^2\right)$
	$\mu_\beta \sim N\left(1, 1000^2\right)$
	$\sigma_\beta^2 \sim IG(0.01, 0.001)$

$IW_{2+1}(.)$: inverse Wishart prior with 2+1 degrees of freedom; $N(.,.)$: normal prior; $IG(.,.)$: inverse gamma prior; $\Gamma(.,.)$: gamma prior; I_2 represents and identity matrix of size 2.

JAGS-code: Speed-Accuracy Model

```
model {
  for (j in 1:N){
    for (i in 1:K){
      # item responses
      U[j, i] ~ dbern(prob[j, i])
      logit(prob[j, i]) <- PersPar[j,1] - ItemPar[i,1]
      # response times
      RT[j,i] ~ dlnorm( muOfLogX[j,i] , alpha.sqr  )
      muOfLogX[j,i] <- ItemPar[i,2]- PersPar[j,2]
    }

    # prior for person parameter
    PersPar[j,1:2] ~ dmnorm(muP, invSigmaP)
  }

  # hyperprior for person parameter
  muP <- c(0,0)
  invSigmaP ~ dwish(M,3)
  SigmaP <- inverse(invSigmaP)
  correlP <- SigmaP[1,2]/(sqrt(SigmaP[1,1])*sqrt(SigmaP[2,2]))

  # prior for item parameter
  # prior for alpha
  alpha.sqr ~ dgamma(0.01, 0.001)
  alpha<-sqrt(alpha.sqr)
  for (i in 1:K){
    ItemPar[i,1:2] ~ dmnorm(muI, omegaI)
  }
  # hyperprior for item parameter
  muI[1] ~ dnorm(0, 0.000001)
  muI[2] ~ dnorm(1, 0.000001)
  invSigmaI ~ dwish(M,3)
  SigmaI <- inverse(invSigmaI)
  correlI <- SigmaI[1,2]/(sqrt(SigmaI[1,1])*sqrt(SigmaI[2,2]))
}
```

Note. N: number of persons, K: number of items. M represents an identity matrix of size 2. U is a N by K matrix containing the item responses and RT is a N by K matrix containing the associated response times.

JAGS-code: Manifest missing data model

```
model {
  for (j in 1:N){
    for (i in 1:K){
      U[j, i] ~ dbern(prob[j, i])
      logit(prob[j, i]) <- theta[j] - b[i]
    }

    # prior for person parameter
    theta[j] ~ dnorm(muP[j], invSigmaP)
    muP[i] <- gamma[1] + gamma[2]*Z[j]
  }

  # prior for item difficulties
  for (i in 1:K) {
    b[i] ~ dnorm(muI, invSigmaI)
  }

  # identification and prior for beta
  gamma[1] <- 0
  gamma[2] ~ dnorm(0,0.000001)

  # hyperprior for person parameter
  invSigmaP   ~ dgamma(0.01, 0.001)
  SigmaP <- 1/invSigmaP

  # hyperprior for item parameter
  muI ~ dnorm(0, 0.000001)
  invSigmaI   ~ dgamma(0.01, 0.001)
  SigmaI <- 1/invSigmaI
}
```

Note. N: number of persons, K: number of items. U is a N by K matrix containing the item responses. Z is a vector of length n representing the number of not-reached items.

Appendix B

Difference in Speed Estimates

(see Figs. 14, 15, 16, 17.)

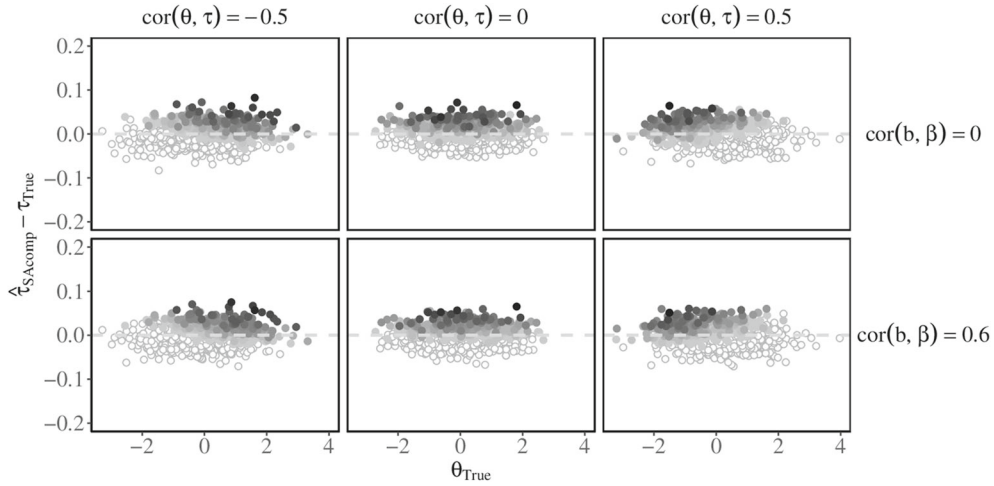

FIGURE 14.
Difference in speed estimates using the SA model for complete data compared to the true speed values as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

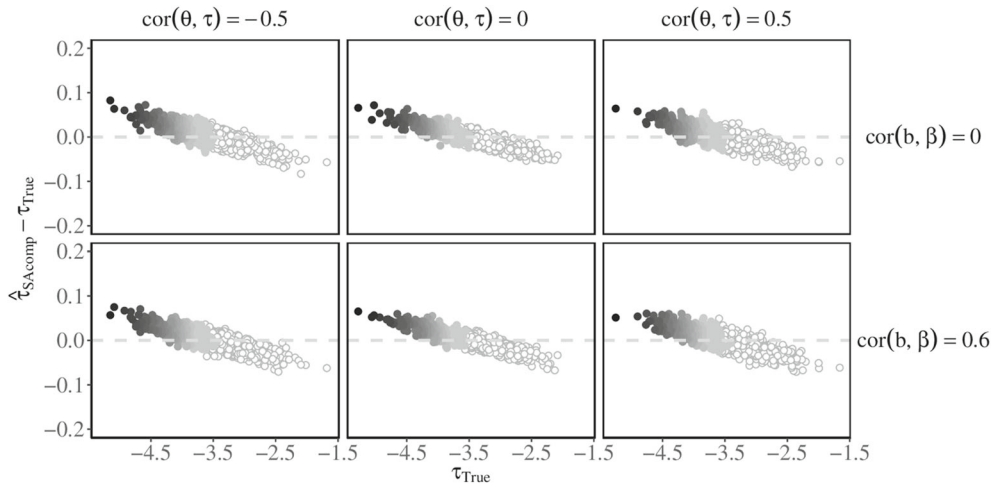

FIGURE 15.
Difference in speed estimates using the SA model for complete data compared to the true speed values as a function of true speed. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

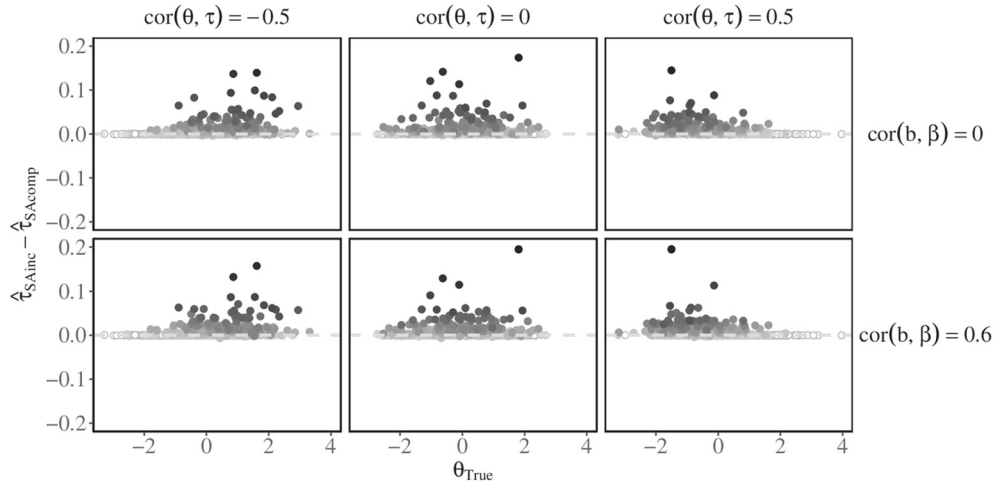

FIGURE 16.
Difference in speed estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

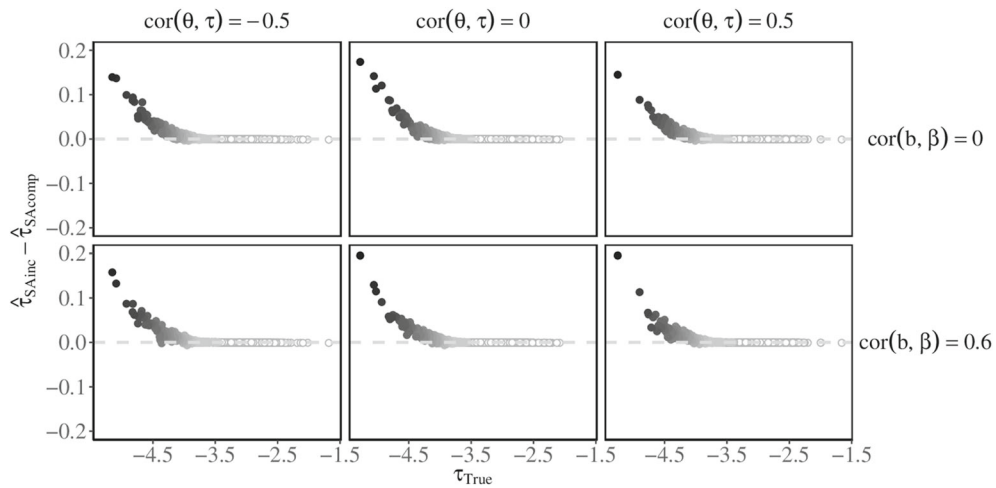

FIGURE 17.
Difference in speed estimates between the SA model for incomplete data (SAinc) and the SA model for complete data (SAcomp) as a function of true speed. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

Appendix C

Subsequent Analyses

(see Figs. 18, 19.)

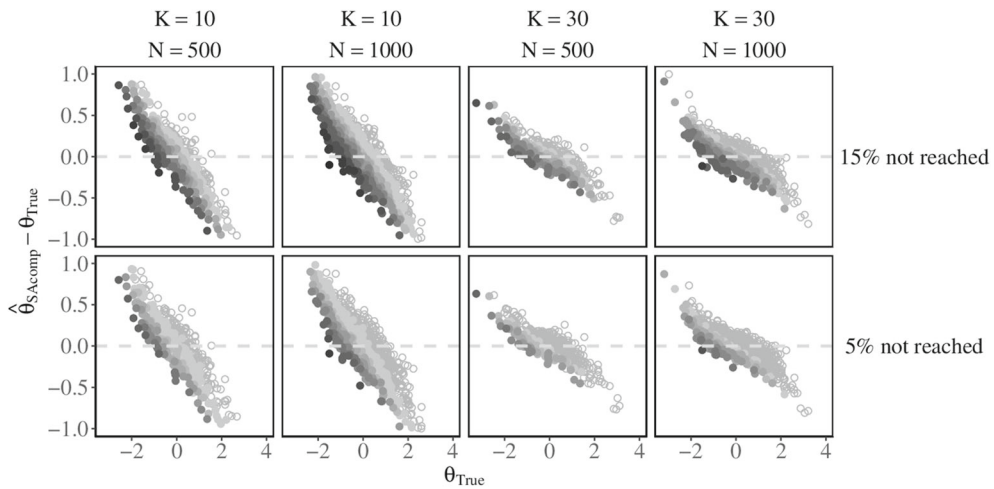

FIGURE 18.
Difference in ability estimates using the SA model for complete data (SAcomp) compared to the true ability values as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color, with darker colors denoting a higher number of not-reached items.

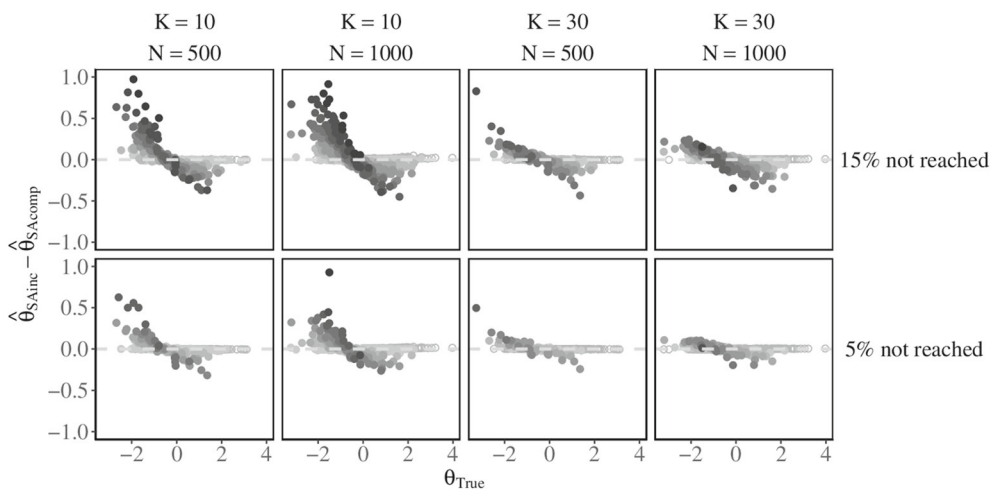

FIGURE 19.
Difference in ability estimates between using the SA model for incomplete data (SAinc) and using the SA model for complete data (SAcomp) as a function of true ability. White circles represent simulees without missing values and filled circles persons with missing values. The number of missing values is given by the circles' color with darker colors denoting a higher number of not-reached items.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report (NCES 2001–509)*. Washington, DC: National Center for Education Statistics.

Bolsinova, M., & Tijmstra, J. (2016). Modeling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148. https://doi.org/10.1007/s11336-016-9537-6.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257–279. https://doi.org/10.1111/bmsp.12076.

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

Culbertson, M. (2011). *Is it wrong? Handling missing responses in IRT*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, 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. *Educational and Psychological Measurement*, *38*, 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, *161*(3), 1307–1320.

Duchhardt, C. & Gerdes, A. (2012). *NEPS technical report for mathematics—scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Berlin: Springer.

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of markov chain monte carlo* (pp. 163–174). London: Chapman and Hall/CRC.

Goegebeur, Y., De Boeck, P., Wollack, J., & Cohen, A. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87. https://doi.org/10.1007/s11336-007-9031-2.

Glas, C. A., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*, *57*(4), 523–541.

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13:3–4*, 133–164. https://doi.org/10.1080/15366367.2015.1100020.

Goldhammer, F., & Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response time modeling. *Applied Psychological Measurement*, *38*(4), 255–267. https://doi.org/10.1177/0146621613517164.

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. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 technical report: (Rep. No. 21-TR-20)*. NJ: Princeton.

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21. https://doi.org/10.1007/s11336-008-9075-y.

Köhler, C., Pohl, S., & Carstensen, C. H. (2014). Taking the missing propensity into account when estimating competence scores: Evaluation of IRT models for non-ignorable omissions. *Educational and Psychological Measurement*, *1*, 1–25. https://doi.org/10.1177/0013164414561785.

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

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. New York: Academic Press.

Kuhn, J.-T., & Ranger, J. (2015). Measuring speed, ability, or motivation: A commentary on Goldhammer (2015). *Measurement: Interdisciplinary Research and Perspectives*, *13:3–4*, 173–176. https://doi.org/10.1080/15366367.2015.1105065.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large Scale Assessments in Education*, *2*(8), 1–24. https://doi.org/10.1186/s40536-014-0008-1.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264. https://doi.org/10.1007/BF02291471.
Meng, X. B., Tao, J., & Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*, 1–27. https://doi.org/10.1111/jedm.12060.
Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*, i–36. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x.
Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626. https://doi.org/10.1080/00273171.2016.1192983.
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219. https://doi.org/10.1111/bmsp.12042.
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*(3), 445–459. https://doi.org/10.1111/1467-985X.00177.
OCED. (2009). *PISA 2006 technical report*. Paris: OECD.
OCED. (2017). *PISA 2015 technical report*. Paris: OECD.
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*(2), 177–194.
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol.124).
Plummer, M. (2016). *rjags: Bayesian graphical models using MCMC. R package version 4-6*. Retrieved from https://CRAN.R-project.org/package=rjags.
Pohl, S. & Carstensen, C. (2012). *NEPS technical report—scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-University, 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*(3), 423–452. https://doi.org/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.
Pohl, S., & von Davier, M. (2018). Commentary: "On the importance of the speed-ability trade-off when dealing with not reached items" by Jesper Tijmstra and Maria Bolsinova. *Frontiers in Psychology*, *9*, 1988. https://doi.org/10.3389/fpsyg.2018.01988.
R Development Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.r-project.org.
Ranger, J., & Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31–47. https://doi.org/10.1007/s11336-011-9231-7.
Ranger, J., & Orthner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128–148.
Rose, N. (2013). *Item nonresponses in educational and psychological measurement (Unpublished doctoral dissertation)*. Friedrich-Schiller-University of Jena, Germany.
Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*, i–53. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x.
Sachse, K., Mahler, N., & Pohl, S. (2019). When nonresponse mechanisms change: effects on trends and group comparisons in international large-scale assessments. *Educational and Psychological Measurement*,. https://doi.org/10.1177/0013164419829196.
Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232.
Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.
Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement*, *35*(6), 433–446. https://doi.org/10.1177/0146621611407305.
Senkbeil, M. & Ihme, J. M. (2012). *NEPS Technical report for computer literacy—scaling results of starting cohort 4 in ninth grade* (NEPS Working Paper No. 17). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.
Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology*, *9*, 964. https://doi.org/10.3389/fpsyg.2018.00964.
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204. https://doi.org/10.3102/10769986031002181.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z.
van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20. https://doi.org/10.3102/1076998607302626.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. https://doi.org/10.1111/j.1745-3984.2007.00030.x.

van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139. https://doi.org/10.1007/s11336-009-9129-9.

van der Linden, W. J., & Guo, F. (2008). Bayesian Procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to dontrol for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. https://doi.org/10.1177/01466219922031329.

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. Special issue: Current methodological issues in large-scale assessments. *Psychological Test and Assessment Modeling*, *58*(4), 671–701.

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

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost (Ed.), *Applications of latent trait nd latent class models in the social sciences* (pp. 89–98). Münster, Germany: Waxmann.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC Cognitive Data. In Organisation for Economic Cooperation and Development (2013), Technical Report of the Survey of Adult Skills (PIAAC) (pp. 406–438). OECD Publishing. Available at: http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf.

Manuscript Received: 25 OCT 2018
Published Online Date: 3 MAY 2019