Educational and Psychological
Measurement
2019, Vol. 79(1) 170–199

DOI: 10.1177/0013164417749679
journals.sagepub.com/home/epm
SAGE

A Semiparametric Approach for Modeling Not-Reached Items

Marit Kristine List[1], Olaf Köller[1], and Gabriel Nagy[1]

Abstract

Tests administered in studies of student achievement often have a certain amount of not-reached items (NRIs). The propensity for NRIs may depend on the proficiency measured by the test and on additional covariates. This article proposes a semiparametric model to study such relationships. Our model extends Glas and Pimentel's item response theory model for NRIs by (1) including a semiparametric representation of the distribution of the onset of NRIs, (2) modeling the relationships of NRIs with proficiency via a flexible multinomial logit regression, and (3) including additional covariates to predict NRIs. We show that Glas and Pimentel's and our model have close connections to event history analysis, thereby making it possible to apply tools developed in this context to the analysis of NRIs. Our model was applied to a timed low-stakes test of mathematics achievement. Our model fitted the data better than Glas and Pimentel's model, and allowed for a more fine-grained assessment of the onset of NRIs. The results of a simulation study showed that our model accurately recovered the relationships of proficiency and covariates with the onset of NRIs, and reduced bias in the estimates of item parameters, proficiency distributions, and covariate effects on proficiency.

Keywords

educational assessment, item response theory, not-reached items, event history analysis, latent class analysis, nonlinear relations

Within scientific studies of student achievement, tests are typically administered with a time limit. As a consequence, students might not reach the end of a test within the allotted testing time, and this can lead to a special type of missing data, reflected in

[1] Leibniz Institute for Science and Mathematics Education at Kiel University, Kiel, Germany

Corresponding Author:
Marit Kristine List, Educational Measurement, Leibniz Institute for Science and Mathematics Education at Kiel University, Olshausenstraße 62, Kiel, 24118, Germany.
Email: list@ipn.uni-kiel.de

the number of not-reached items (NRIs). NRIs imply a monotone pattern of missing data; that is, all items located after the first item not reached are missing. Hence, the earlier the onset of NRIs, the more item responses are missing in a test. NRIs have received some attention in the psychometric literature because the onset of NRIs appears to be related to the proficiency measured by the test (Glas & Pimentel, 2008; Köhler, Pohl, & Carstensen, 2015a; Lawrence, 1993; Pohl, Gräfe, & Rose, 2014). As a reaction to the problem, item response theory (IRT; Embretson & Reise, 2000) models that make it possible to estimate the relationships between proficiency and the onset point of NRIs have been developed, with the approach suggested by Glas and Pimentel (2008) being most frequently used (e.g., Pohl et al., 2014). However, the model of Glas and Pimentel (2008), as well as many other approaches (e.g., Hutchison & Yeshanew, 2009), assumes a linear relationship between proficiency and the onset of NRIs: an assumption that might be questioned in many applications.

Linear relationships between NRIs and the proficiency measured by the test appear most plausible in testing situations in which NRIs can be conceived as pure reflections of test speededness (e.g., Evans & Reilly, 1972), such as in the case of high-stakes tests, where test takers invest full effort to respond to all items in the test. Here, students with higher proficiencies might solve the items at a higher pace, which means that they are more likely to reach the end of the test. However, even in situations in which test takers show their maximum performance, nonlinear relationships between proficiency and NRIs could exist because test takers at low proficiency levels might reach the end of the test by applying too simplistic or quick strategies to difficult items. In this case, test takers at an intermediate proficiency level might then show the slowest solution behavior, leading to an earlier onset of NRIs. Relationships between proficiency and the onset point of NRIs appear likely to be complex in situations in which students are not motivated to show their maximum performance. In low-stakes assessments, test-taking behavior has been found to be related to test-taking motivation (Wise & DeMars, 2005), which means that NRIs could be affected by motivational reactions to the test. Therefore, students with low proficiencies might either have an early onset of NRIs because they become frustrated with the test, or they might complete the test without investing much effort. Hence, the distribution of NRI onsets could be multimodal for certain levels of proficiency.

In addition, the amount of NRIs could also depend on person characteristics as well as on the proficiency being measured by the test (e.g., Dorans, Schmitt, & Bleistein, 1992; Evans & Reilly, 1972; Köhler et al., 2015a; Schmitt, Dorans, Crone, & Maneckshana, 1991). In principle, person variables could be related to the students' proficiencies and to the onset of NRIs. Hence, the question arises of whether NRIs can be fully predicted by the proficiency variable, or whether the person covariates have an additional impact on NRIs when proficiency is held constant. This question has some similarity with the concept of *differential test functioning* (Shealy & Stout, 1993) that refers to the question of whether a test works differently for examinees with the same proficiency but taken from different groups. Therefore, a finding that a covariate has an effect on NRIs while conditioning on proficiency indicates a *differential onset of NRIs* in examinees of the same proficiency, implying a systematic difference in the amount of information provided (i.e., the number of item responses preceding NRIs) to measure proficiency at the different levels of the covariate. Such differences could be due to different mechanisms, such as mastery in the test language or test-taking motivation, among others. Native speakers have been found to respond to more items in the allotted testing time (Schmitt & Bleistein, 1987; Sireci, Han, & Wells, 2008), and higher test-taking motivation has been found to be related to more time being spent on tasks (Scherer, Greiff, & Hautamäki, 2015), which could lead to an earlier onset of NRIs. However, as the background characteristics examined could be related to proficiency, in both examples, a rigorous test of the effects of students' background characteristics on NRIs requires the impact of proficiency on NRIs to be accounted for.

The aim of the present article is to provide a flexible and easy-to-use IRT approach for modeling the onset of NRIs as a possibly nonlinear function of the proficiency measured by the test, as well as of additional person covariates. Our model combines a two-parameter logistic (2PL) IRT model (Birnbaum, 1968), applied to the item responses, with a latent class model (LCM; Formann, 1985), applied to the indicators of NRIs. Our LCM can be conceived as a semiparametric version of the continuous steps model for assessing the onset of NRIs suggested by Glas and Pimentel (2008). In our approach, the relationships of the onset of NRIs with the proficiency variable and the additional covariates were modeled via a multinomial logit regression, thereby allowing for nonlinear relationships. The newly proposed model was applied to a timed low-stakes test of mathematics achievement in order to demonstrate its utility in applied settings. In a small simulation study, we further investigated whether the model correctly recovers the relationships of NRIs with proficiency and covariates, and whether our approach reduces biases in the estimates of item parameters, proficiency distributions, and covariate effects on proficiency that are often found in IRT models that disregard missing responses (e.g., Rose, von Davier, & Xu, 2010).

Relationships of NRIs With Proficiency and Covariates

Several studies have documented relationships between the onset of NRIs and the characteristics of test takers. Results suggest that the amount of NRIs is higher in ethnic minority groups (Dorans et al., 1992; Schmitt & Bleistein, 1987; Schmitt et al., 1991) but does not appear to differ between gender groups (Evans & Reilly, 1972; Schmitt et al., 1991; Wild, Durso, & Rubin, 1982). Some more recent studies have examined the relationships between NRIs and the proficiency measured by the test by adopting the IRT approach suggested by Glas and Pimentel (2008). These studies provide evidence for statistically significant relationships between proficiency and the onset of NRIs, but the pattern of results differed between tests and samples (Glas & Pimentel, 2008; Pohl et al., 2014). Most recently, Köhler et al. (2015a) studied the predictors of NRIs in reading tests implemented in several age groups. Their analyses revealed reading speed to be a strong and consistent predictor of NRIs, in that faster readers had a later onset of NRIs. Köhler et al. (2015a) employed the number of NRIs as a dependent variable in linear regression analyses. Similarly, researchers employing the model of Glas and Pimentel (2008) also did not investigate the nonlinear relationships that, as we have described above, appear plausible in the case of NRIs.

Investigating the effects of covariates on NRIs while simultaneously controlling for latent proficiency might be of interest for two reasons. First, in studies aiming to describe the distribution of student proficiencies in different subpopulations, the differential onset of NRIs indicates a threat to the validity of group comparisons as it means that groups differ in their test-taking behavior. Thus, group differences in proficiency might be different if respondents of the same proficiency level show the same test-taking behavior. Second, the effect of covariates on the onset of NRIs, while controlling for proficiency, could be a key research question in some applications. For example, researchers might hypothesize that a specific curricular intervention raises students' proficiencies and, in addition, enhances the pace at which students work on the test, thereby reducing the number of NRIs. To provide support for this hypothesis, a result indicating a differential onset of NRIs would be required.

Taken together, the differential onset of NRIs indicates that proficiency does not provide a sufficient explanation for differences in the onset of NRIs across different levels of a covariate. Therefore, differential onsets of NRIs indicate that test takers that differ with respect to the covariate's value, but not to the level of proficiency, show different test-taking behavior. The differential onset of NRIs could indicate that the equivalence of measuring the proficiency variable across groups is violated. However, whether such a finding is considered as a threat to the validity clearly depends of the aims of the investigation.

IRT Models for Missing Responses and the Onsets of NRIs

Missing item responses in tests are regarded as problematic because they are likely to be related to the proficiencies being measured. As such, the missing data are nonignorable (NMAR; Little & Rubin, 2002), which means that missing data mechanisms need to be included in the model in order to prevent biased parameter estimates. To accomplish this task within the framework of the IRT, the full data likelihood that includes the vector of item scores $\boldsymbol{Y}$, a set of missing-data indicators $\boldsymbol{D}$, and possibly a set of covariates $\boldsymbol{X}$, that is $P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X})$, needs to be considered. Consideration of $\boldsymbol{X}$ makes it possible to investigate whether the relationships between $\boldsymbol{Y}$ and $\boldsymbol{D}$ vanish once accounting for $\boldsymbol{X}$, so that the missing-data process is turned into a missing-at-random (MAR; Little & Rubin, 2002) process. Under MAR, $\boldsymbol{D}$ no longer contributes to the estimation of parameters that apply to $\boldsymbol{Y}$, and can therefore be ignored (e.g., Glas, Pimentel, & Lamers, 2015). Different types of modeling strategies have been applied to $P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X})$. In typical IRT applications, the item scores $\boldsymbol{Y}$ are assumed to reflect one or multiple continuous proficiency variables, so that the relationships between $\boldsymbol{Y}$ and $\boldsymbol{D}$ are modeled via the relationships of the proficiency variables with $\boldsymbol{D}$ (e.g., Rose, von Davier, & Nagengast, 2015).

Pattern mixture IRT models (Little, 1994) aim to stratify the sample according to distinct missing-data patterns. They provide indications of NMAR patterns when the proficiency variable (e.g., its mean) differs between strata. Practically, these models can be implemented either by means of multigroup IRT models, in which the groups are defined by distinct patterns of missingness, or by regressing the proficiency variables on indicators of the missing-data patterns, as well as on the covariates (e.g., Rose, von Davier, & Nagengast, 2017). In the context of NRIs, Rose et al. (2010) suggested regressing the continuous proficiency variable on the number of the individuals' NRIs. The model can be extended to include $\boldsymbol{X}$, as well as nonlinear relationships between proficiency and NRIs, for example, by using the polynomial functions of the amount of NRIs. Although the model is quite flexible and easy to use, it is not well suited for the purpose of studying the determinants of NRIs, because $\boldsymbol{D}$ is treated as an independent variable.

A second type of models assumes that additional latent variables underlie the missing-data indicators $\boldsymbol{D}$, which means that the relationships between proficiency and $\boldsymbol{D}$ are modeled via the relationships between latent variables. Most often, the joint distribution of latent variables is assumed to be multivariate and normal. These kinds of IRT models can be considered to belong to the family of *shared parameter models* (Wu & Carroll, 1988). They have been extended to multidimensional representations of $\boldsymbol{D}$, with the possibility of combining indicators of omissions with indicators of NRIs (Rose et al., 2017), and of including the covariates $\boldsymbol{X}$ affecting all latent variables (Glas, Pimentel, & Lamers, 2015). A drawback of these models is that the (conditional) multivariate normality assumption implies linear relationships between latent variables; this might be called into question. Köhler, Pohl, and Carstensen (2015b) relaxed the linearity assumption by formulating a two-dimensional IRT model for proficiency and omitted responses as a general diagnostic model (GDM; von Davier, 2008). GDMs allow any kind of multivariate distribution to be approximated by discretizing the latent variables into different prespecified levels. However, to the best of our knowledge, GDMs for missing responses have not yet been extended to include continuous covariates, and have not been applied to NRIs.

Selection models (Little & Rubin, 2002) refer to the third type of models that can be applied to account for NMAR patterns. Here, the full data likelihood $P(\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{X})$ is factorized into the distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$, $P(\boldsymbol{Y}|\boldsymbol{X})$, and the probability of missing data $\boldsymbol{D}$ given $\boldsymbol{Y}$ and $\boldsymbol{X}$, $P(\boldsymbol{D}|\boldsymbol{Y},\boldsymbol{X})$, such that (e.g., Rose et al., 2017):

$$P(\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{X}) = P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{D}|\boldsymbol{Y},\boldsymbol{X}). \tag{1}$$

Because $\boldsymbol{D}$ is expressed as an outcome variable, selection IRT models provide a natural way for studying the determinants of missing responses, including NRIs. Within the IRT there are different ways of specifying models in which $\boldsymbol{D}$ depends on proficiency. In some models, it is assumed that the dependencies between all indicators $\boldsymbol{Y}$ and $\boldsymbol{D}$ can be fully explained by the proficiency variables. This assumption has been relaxed in other applications by including an additional continuous latent variable, so that $\boldsymbol{D}$ is simultaneously affected by multiple dimensions. In addition, the models can be extended to include covariates assumed to affect proficiencies, as well as the missing-data indicators. Furthermore, Bacci and Bartolucci (2015) relaxed distributional assumptions in these models by discretizing the latent variable. Their model is very flexible because it allows the relationships of proficiencies and covariates with missing responses to be item-specific. However, its drawback is that it includes many parameters that might not be reliably estimated in the case of a small percentage of missing values. Therefore, examining the determinants of the onsets of NRIs in the context of selection IRT models calls for an approach that consists of a suitable number of parameters to be estimated, and that allows for nonlinear effects of all the variables, including proficiencies. Furthermore, the model should allow for a flexible assessment of relationships.

NRIs reflect a special kind of missing data, because once an item response is missing in the sequence of test items, all responses that follow the first missing response are also missing. Hence, this pattern of missing data can be regarded as being irreversible. Such situations are often at the core of longitudinal investigations that focus their attention on the risk (or hazard) that some irreversible events occur over time by employing methods known as event history analysis or survival analysis (Allison, 2014; Singer & Willett, 1993). Hence, the occurrence of NRIs over the sequence of test items, as represented by $P(\boldsymbol{D}|\boldsymbol{Y},\boldsymbol{X})$ in Equation (1), can be examined by similar methods, with the difference that the (discrete) time points are replaced by discrete positions in a sequence of items. As we will show in the next section, when reformulated as a selection model, the model proposed by Glas and Pimentel (2008) can be regarded as a type of discrete-time event history model.

Glas and Pimentel's (2008) Model for NRIs in Speeded Tests

Glas and Pimentel (2008) considered NRIs and proposed a two-dimensional IRT model that includes a latent proficiency dimension and a second dimension indicating the number of items attempted by the examinees (i.e., a steps variable). The indicators of the proficiency variable are the actual item responses. In the case of dichotomous item responses, the proficiency variable is defined according to the 2PL model such that:

$$logit\left[P\left(y_{ij}=1|\theta_i\right)\right]=\alpha_j\left(\theta_i-\beta_j\right), \tag{2}$$

which means that the logit of the probability of a correct item response of individual i (i = 1, 2,. . ., N) to item j (j = 1, 2, . . ., J), y_{ij}, is a function of the individual's proficiency θ_i. In Equation (2), α_j and β_j stand for the discrimination and difficulty of item j.

The latent variable that assesses the onset of NRIs is measured by means of *response indicators* that are defined as follows: For each examinee, the vector of response indicators consists of a series of "1" for all items to which the examinee responds, followed by, at most, one ‘‘0’’ for the first NRI, and *missing flags* for all subsequent NRIs. For example, in a hypothetical seven-item test, an examinee who does not reach the last three items receives a vector of response indicators of $\boldsymbol{d}_i'$= [1111099], where ‘‘9’’ indicates a missing value.

The steps variable underlying response indicators is defined by the *steps model* (Verhelst, Glas, & de Vries, 1997), which expresses the probability that a response is observed in a particular item position, given that all former item responses were observed. Glas and Pimentel (2008) presented applications in which the steps model included only a difficulty parameter, similar to the Rasch model (Rasch, 1960), such that

$$logit\left[P\left(d_{ij}=1|\xi_i\right)\right]=\xi_i-\tau_j, \tag{3}$$

where d_{ij} is the value of the response indicator for person i for item position j, ξ_i stands for the examinee i’s steps variable with zero mean and unconstrained variance, and τ_j is a difficulty parameter of the response indicator in item position j. In the steps model, the difficulty parameters τ_j are constrained to follow a linear function across item positions, $\tau_j=t_0+(j-J)t_1$, which means that only two parameters are estimated (i.e., t_0 and t_1). The latent variable ξ measures the number of steps taken by an examinee (i.e., number of items that are not NRIs): The lower the values of the steps variable ξ, the earlier the onset of NRIs.

The proficiency variable θ and the steps variable ξ are assumed to have a bivariate normal distribution with the correlation coefficient $\rho_{\xi,\theta}$. Positive correlations indicate that higher proficiencies are associated with later onsets of NRIs, whereas negative relationships indicate that higher proficiencies are related to earlier onsets of NRIs. In applications of Glas and Pimentel’s (2008) model, the absolute size of the correlation coefficient is regarded as an indicator of NMAR missing-data patterns. However, as our focus is on the relationship between proficiency and the onset of NRIs, we focused on a reparametrized version of their model in which ξ was treated as a dependent variable predicted by θ.

Relationships to Discrete-Time Event History Analysis. As previously mentioned, ξ can be treated as a variable depending on θ, such that

$$\xi_i=\gamma_{\xi,\theta}\theta_i+\zeta_{\xi,i}, \tag{4}$$

with $\gamma_{\xi,\theta}$ being a structural regression weight, and ζ_ξ standing for a normally distributed residual with zero mean and unconstrained variance that is uncorrelated with θ. By combining Equations (3) and (4), the relationships of θ with the response indicators d_j can be represented as

$$logit\left[P\left(d_{ij}=1|\theta_i,\zeta_{\xi,i}\right)\right]=\gamma_{\xi,\theta}\theta_i+\zeta_{\xi,i}-\tau_j, \tag{5}$$

showing that the model implies the same effect of the proficiency variable on the logit of each response indicator irrespective of its position.

Equation (5) has direct connections to discrete-time event history analysis, which models the effect of a variable on the probability that an irreversible event will occur, given that the event has not occurred before. If the ideas of event history analysis are applied to the phenomenon of NRIs, the focus is on $P(d_{ij}=0)$, marking the probability of not making the step from item $j-1$ to item j (given that all previous steps were taken), instead of on $P(d_{ij}=1)$, which refers to the probability of making the step from item $j-1$ to item j. To derive the hazard probability, the right-hand side of Equation (5) could be multiplied by -1, such that

$$logit\left[P\left(d_{ij}=0|\theta_i,\zeta_{\xi,i}\right)\right]=\tau_j-\left(\gamma_{\xi,\theta}\theta_i+\zeta_{\xi,i}\right). \tag{6}$$

Hence, the model of Glas and Pimentel (2008) can be reformulated as a kind of discrete-time event history model that is applied to the sequence of test items instead of to the sequence of time points, which means that the model can be understood as a *discrete (item) sequence event model* (DSEM). In Equation (6), θ serves as an explanatory variable, and the person variable ζ_ξ reflects a *frailty factor* (Allison, 2014; B. Muthén & Masyn, 2005) that accounts for heterogeneity in the hazards of NRI onsets not explained by θ. Similar to traditional discrete-time event models (Allison, 2014), the DSEM representation of the model of Glas and Pimentel (2008) assumes that the logits of all items' hazard probabilities are equally affected by θ because $\gamma_{\xi,\theta}$ is not allowed to vary across items. In event history models, this assumption is known as the *proportional hazards assumption*, which means that the logit-hazard profiles (defined by j = 1, 2, . . ., J) predicted by θ are proportional to one another (i.e., they have a common shape and are mutually parallel; Singer & Willett, 1993). Note that the DSEM presented in Equation (6) is more restrictive than conventional event history models because the parameters τ_j are constrained to follow a linear function, thereby forcing the baseline hazards of NRIs to increase over the course of the test. Typical discrete-time event models do not incorporate such assumptions and leave the τ-parameters unconstrained.

The DSEM formulation of Glas and Pimentel's (2008) model makes it possible to apply the graphical tools developed in the context of event history analysis to the onset of NRIs. Here, we focus on the *survival function*, which depicts the probability of "surviving" over a sequence of m ($m \leq J$) items as a function of the explanatory variable θ. In the context of the present model, the survival function can be written as

$$P\left(S_i \geq m|\theta_i,\zeta_{\xi,i}\right)=\prod_{j=1}^{m}P\left(d_{ij}=1|\theta_i,\zeta_{\xi,i}\right), \tag{7}$$

where $S=1,2,\ldots,J$ denotes the individual survival variable, whose value is equal to the last item a person has responded to. The survival function allows for a compact representation of the relationships between person variables and the probability of completing the test up to a specific point. In real applications, the survival function $P(S_i \geq m|\theta_i)$, defined over the full distribution of the frailty factor, might be more relevant than the survival function specified for specific combinations of θ and ζ_ξ (Equation 7). Deriving $P(S_i \geq m|\theta_i)$ requires determining the average of the function given in Equation (7) over the full distribution of ζ_ξ, which might turn out to be quite cumbersome in practice. This goal can be achieved by integrating over $\zeta_{\xi,i}$, or by simulating the distribution survival functions at the desired values of θ.

To sum up, the model suggested by Glas and Pimentel (2008) can be reformulated as a DSEM. This shows that Glas and Pimentel's (2008) model builds upon the proportional hazards assumption, that is, it specifies that θ has an equal impact on all response indicators. Their model also makes strong assumptions about the increasing baseline hazards (for a discussion, see Pohl et al., 2014). Furthermore, the model builds upon the assumption of a normally distributed frailty factor.

A Semiparametric Model for the Onset of NRIs

In this section, we present a semiparametric version of the model of Glas and Pimentel (2008), which makes less strong assumptions about the distribution of the steps variable (i.e., the ξ-variable in Equation 4) and relaxes the proportional hazards assumption. In addition, we extend the model to include covariates, so that the model allows the differential onset of NRIs to be examined. Our approach builds upon the (continuous) 2PL model for item responses as represented in Equation (2), and on a semiparametric representation of the steps model given in Equation (4). To provide a metric for the steps variable that can be interpreted more easily, we propose a different parameterization of Equation (3). We define a new steps variable δ that can be understood as a direct measure of the number of steps taken by an individual:

$$logit\left[P\left(d_{ij}=1|\delta_i\right)\right]=\lambda(\delta_i-j). \quad (8)$$

In Equation (8), λ stands for a discrimination parameter that applies to all response indicators. The difficulty parameters given in Equation (3) are replaced by the item indexes $j = 1, 2, \ldots, J$. Because the difficulties in Equation (8) are fixed, the mean of δ can now be estimated. The latent variable δ is measured in units of item positions, such that $\delta_i = j$ provides a probability of .5 of providing an item response in position j. The probabilities of providing responses to items preceding j are higher than .5 and the probabilities for item positions after j are lower than .5. Because the items are assumed to be equally spaced, the probability curve follows a logistic function. The steepness of the function is driven by λ, with higher values indicating more abrupt changes in the probability of responding around the individual inflection point δ_i.

In our semiparametric version of the steps model, we specify δ to have a discrete distribution, with K support points δ_k ($k = 1, 2, \ldots, K$) that are freely estimated, in order to derive a flexible representation of the distribution of the δ-variable. This approach is equivalent to a *located latent class model* (Formann, 1985), which includes a latent class variable c with K categories and class proportions π_k with $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The purpose of our LCM application was to relax the assumption of a normally distributed steps variable (and, therefore, of a normally distributed frailty factor), and to approximate the unknown distribution of δ (B. Muthén & Masyn, 2005; Masyn, 2009). Our approach makes it possible to relax the linearity assumption included in the model of Glas and Pimentel (2008), which implies the proportional hazards assumption. Our approach can be considered as an *indirect application* of an LCM (Bauer, 2005), where the primary task is not to classify individuals, but to relax distributional and functional assumptions. The LCM part can be represented by altering Equation (8) to

$$logit\left[P\left(d_{ij}=1|c_i=k\right)\right]=\lambda(\delta_k-j) \tag{9}$$

As shown in Equation (9), δ is assumed to be constant within latent classes, which means that, for each class, only one pattern of probabilities exists for all J response indicators. The values of δ_k can be considered as the *support points* of a nonparametric approximation of an arbitrary continuous distribution, whereas the latent class proportions π_k can be considered as the weights associated with support points (Aitkin, 1999; Bacci & Bartolucci, 2015).

The model for response indicators has connections to mixture discrete-time event history analysis (B. Muthén & Masyn, 2005). As the pattern of probabilities of response indicators is constant within each latent class, each class provides one survival function:

$$P(S_i \geq m|c_i=k)=\prod_{j=1}^{m} P\left(d_{ij}=1|c_i=k\right). \tag{10}$$

Furthermore, the marginal survival function can be derived by summing over the class-specific survival functions (Equation 10) weighted by their class proportions:

$$P(S_i \geq m)=\sum_{k=1}^{K} \pi_k P(S_i \geq m|c=k). \tag{11}$$

The semiparametric approach allows for heterogeneity in the latent steps variable δ, since the LCM approach circumvents the assumption of a normally distributed latent steps variable.

The main motivation behind our semiparametric approach was to relax the proportional hazards assumption implicitly made in the model of Glas and Pimentel (2008; see Equation 6). Here, we estimate the impact of θ on δ by means of multinomial regression, relating the latent class variable c to θ, such that

$$P(c_i=l|\theta_i)=\frac{exp\left(\nu_l+\gamma_{l,\theta}\theta_i\right)}{\sum_{k=1}^{K} exp\left(\nu_k+\gamma_{k,\theta}\theta_i\right)}, \tag{12}$$

where ν_l is the structural intercept specific to class c = l, and $\gamma_{l,\theta}$ is the regression weight relating the proficiency variable θ to the latent class c = l. We impose the typical identification restriction of fixing the multinomial parameters of the class $c = K$ to zero.

By using the multinomial regression (Equation 12), the proportional hazards assumption is relaxed. Since, for each latent class c, a regression coefficient is estimated, an overall nonlinear relationship between proficiency and the onset of NRIs is modeled. For example, the analysis could uncover that θ is only related to membership in classes representing later onsets of NRIs, whereas the continuous approach specifies that θ shows an equal impact on all response indicators. Alternatively, findings could indicate that low levels of proficiency are simultaneously related to membership in classes representing an early onset of NRIs and classes representing a late onset of NRIs.

The model described in Equations (9) and (12) fits into the framework of mixture discrete-time event history models (B. Muthén & Masyn, 2005). Therefore, we refer to this model as *mixture discrete (item) sequence event model* (MDSEM). In the MDSEM, the survival function that depends on θ is derived by summing over the class-specific survival functions (Equation 10) weighted by the class probability predicted by θ (Equation 12):

$$P(S_i \geq m|\theta_i) = \sum_{k=1}^{K} P(c_i = k|\theta_i)P(S_i \geq m|c_i = k). \tag{13}$$

The MDSEM provides a flexible and easy-to-use framework for deriving the survival functions at fixed values of θ. Deriving the survival functions for fixed levels of θ no longer requires integration over a continuous frailty factor, as is the case in the DSEM (Equation 7).

Introducing Additional Covariates. We now address the case of introducing covariates into the model. To keep the presentation simple, we focus on the case with a single covariate x, but note that multiple covariates could be included simultaneously. We begin by using x for predicting θ. Here, we assume a linear relationship, such that

$$\theta_i = \kappa_\theta + \gamma_{\theta,x}x_i + \zeta_{\theta,i} \tag{14}$$

where κ_θ is a structural intercept, $\gamma_{\theta,x}$ is the regression weight of x predicting θ, and ζ_θ is a normally distributed residual term with an expectation value of zero that is assumed to be uncorrelated with all other variables in the model.

The relationship between latent classes and predictors θ and x is now given as

$$P(c_i = l|\theta_i, x_i) = \frac{exp(\nu_l + \gamma_{l,\theta}\theta_i + \gamma_{l,x}x_i)}{\sum_{k=1}^{K} exp(\nu_k + \gamma_{k,\theta}\theta_i + \gamma_{k,x}x_i)}, \tag{15}$$

where the parameters are as defined as before (Equation 12) and subjected to the same identification constraints. The only difference is that the multinomial regression part is extended by the covariate x and its class-specific multinomial regression weight $\gamma_{l,x}$.

Equation (15) is of key importance to the suggested model. More specifically, the hypothesis that missing data caused by NRIs do not depend on the covariate x (or on a set of covariates $\boldsymbol{X}$) implies that $\gamma_{1,x} = \gamma_{2,x} = \cdots = \gamma_{K,x} = 0$, assuming that the covariate's effects on class membership are transmitted via its impact on the proficiency variable (Equation 14). When the model is estimated by maximum likelihood, this hypothesis can be evaluated by means of a likelihood ratio test (LRT) that compares the data likelihood of a full model with a nested model in which the relationships with the latent class variable c are accessed via Equation 12, that is, by setting $\gamma_{1,x} = \gamma_{2,x} = \cdots = \gamma_{K,x} = 0$. A statistically significant LRT provides evidence for the differential onset of NRIs, depending on the covariate considered.

If the covariate x is found to predict latent class membership, its impact can be visualized by using the survival function, evaluated with selected combinations of θ and x:

$$P(S_i \geq m|\theta_i, x_i) = \sum_{k=1}^{K} P(c_i = k|\theta_i, x_i) P(S_i \geq m|c_i = k). \tag{16}$$

Model Estimation and Implementation. The MDSEM can be estimated by maximum likelihood estimation. Indeed, Guo, Wall, and Amemiya (2006) have outlined the estimation of a general class of models, of which the MDSEM is a special case. The joint distribution of the item responses $\boldsymbol{Y}$ and NRI indicators $\boldsymbol{D}$ can be written as (by dropping the symbolic representation of model parameters)

$$P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X}) = \sum_{k=1}^{K} \int P(\boldsymbol{D}|c_i = k) P(\boldsymbol{Y}|\theta) P(c_i = k|\theta, \boldsymbol{X}) P(\theta|X) d\theta, \tag{17}$$

with the full data likelihood function L given by

$$L = \prod_{i=1}^{N} P(\boldsymbol{Y}_i, \boldsymbol{D}_i|\boldsymbol{X}_i) = \prod_{i=1}^{N} \int P(\boldsymbol{Y}_i, \boldsymbol{D}_i, c_i, \theta_i|\boldsymbol{X}_i) d(c_i, \theta_i), \tag{18}$$

whereby the integral includes the continuous integral over θ, as well as summation over c.

Guo et al. (2006) have shown that the model parameters can be estimated by means of the expectation maximization (EM) algorithm, as well as by a Gaussian quadrature with a quasi-Newton algorithm. Hence, the MDSEM can be estimated with different computer programs, including Latent Gold (Vermunt & Magidson, 2005) and M*plus* (L. K. Muthén & Muthén, 1998-2012). In the present article, we used M*plus*, which combines the aforementioned algorithms to a so-called accelerated EM algorithm (EMA). Model estimation starts with the EM algorithm but changes to the quasi-Newton algorithm if EM becomes slow.

There are several issues that need to be considered in practice. The first issue is how to define the metric of the proficiency variable. In most IRT applications, this issue is resolved by standardizing the distribution of θ (i.e., $M = 0$, $SD = 1$). Since, in the general case, θ is specified as an endogenous variable that is impacted by covariates, we suggest freely estimating the (residual) variance term, fixing the mean or the structural intercept of θ to zero (Equation 14), and constraining the discrimination parameters such that they are, on average, one. The latter constraint allows the (residual) variance of the proficiency variable to be freely estimated and does not alter its units of measurement when including covariates to predict θ.

The second issue is that NRIs might not exist in the first positions of a test. As a consequence, all response indicators gathered before the first onset of NRIs have a constant value across all respondents, which means that they should be disregarded in the process of model estimation.

The third issue pertains to the optimal number of latent classes. In applications of latent class analysis, the decision concerning the number of classes to use is typically based on measures of goodness of fit, such as Schwarz's (1978) Bayesian information criterion (BIC; Nylund, Asparouhov, & Muthen, 2007). Users need to be aware that the optimal number of latent classes could also depend on the covariates used for predicting latent class membership (Lubke & B. Muthén, 2005). We suggest basing the decision about the number of classes K on the full MDSEM and keeping K constant across different versions of the model (e.g., models where the covariates are excluded) to make sure that results are not affected by the use of different numbers of classes. Some researchers have suggested identifying the number of latent classes prior to the inclusion of covariates (Kim, Vermunt, Bakk, Jaki, & Van Horn, 2016; Nylund-Gibson & Masyn, 2016). This approach is useful in *direct applications* of the LCM (Bauer, 2005) that require the categorical latent variable to exist independent of the covariates included because individuals' class membership is substantively interpreted. However, the MDSEM is based on an indirect application of the LCM that does not aim to categorize individuals, but only to relax parametric assumptions.

Finally, one problem with maximum likelihood estimation for mixture IRT models is that the solution can converge to a local rather than the global maximum (Finch & French, 2012). Therefore, the usage of multiple random starting values is recommended to ensure replication of the best likelihood value (Lubke & B. Muthén, 2005).

Empirical Illustration

In the next sections, we report on our application of the proposed MDSEM to a timed low-stakes mathematics test taken from a typical large-scale study. This application served three purposes. First, we compared the MDSEM to the model suggested by Glas and Pimentel (2008), thereby demonstrating the flexibility gained by implementing their model in a semiparametric framework. Second, we exemplify how to use the MDSEM to evaluate a test for the differential onset of NRIs while holding the proficiency variable constant. Third, we exemplify the use of graphical procedures to aid the interpretation of model results, while focusing on the survival function.

The models were implemented in M*plus* 7.4 (L. K. Muthén & Muthén, 1998-2012) by using the EMA algorithm using standard integration with 15 integration points for the proficiency variable. All models were estimated using multiple sets of random starts. In all cases, the best log-likelihood was replicated. In order to determine the number of classes, we estimated a series of models ranging from 3 to 8 latent classes. The decision concerning the number of classes was based on the BIC.

Sample and Procedure

The sample was taken from the study ''Mathematics and Science Competence in Vocational Education and Training'' (ManKobE; e.g., Retelsdorf, Lindner, Nickolaus, Winther, & Köller, 2013). It encompassed apprentices in their first year of vocational education and training (VET) in mathematics and science-related occupations, namely, industrial clerks and different technical professions (e.g. industrial and laboratory technicians; further referred to as *technicians*). The test was designed to assess mathematical skills at the core of VET for industrial clerks. The test contained only tasks that could, in principle, be solved with the mathematical knowledge acquired in regular schooling, but the problems presented were embedded in an organizational context typical for industrial clerks. Hence, in this analysis, we expected that industrial clerks would have higher proficiency, and we hypothesized that they would show a later onset of NRIs than technicians because the context in which the items were presented was more familiar to clerks.

We considered the data of $N = 967$ apprentices at the beginning of their VET (average time in VET of about 3 months); cases with less than three valid responses in the whole test and those with missing information on the covariate considered were excluded. From all test takers, $n = 214$ cases were in VET for industrial clerks; the remaining apprentices were in VET for technicians ($n = 753$). On average, apprentices were 18.70 years old ($SD = 2.88$). The test considered consists of 20 dichotomously scored items and it was administered with a time limit of approximately 15 minutes. NRIs were first observed in item position $j = 5$. Therefore, response indicators for the first four items were not included in the analysis.

Results

Descriptive Results. Only 28% of the sample completed the first three quarters of the test and only 16% reached the last item. The sample-based baseline hazard function of the onsets of NRIs is depicted in Figure 1. It appeared that the hazard rates for NRIs did not constantly increase across item positions but rather reached a maximum after the first three quarters of the test (i.e., in position 15). In addition, the hazard function given in Figure 1 appeared to be constituted of several peaks. This led us to expect that the semiparametric steps model was likely to identify several latent classes that are sharply separated from each other.

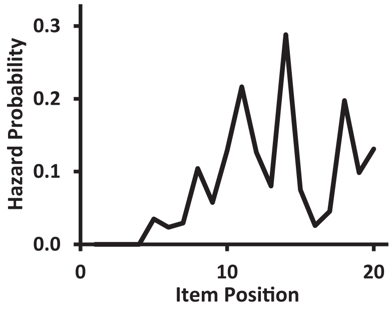

Figure 1. Sample-estimated hazard probabilities of onsets of not-reached items (NRIs).

Nonparametric Representation of the Distribution of Steps Variable. As shown in Table 1, the model with $K = 6$ classes achieved the best fit in terms of the BIC, and we therefore decided on six classes. In this model, the discrimination parameter of response indicators was estimated to be $\hat{\lambda} = 2.16$ ($SE = 0.116$). Hence, a relatively sharp step in the probability of responding to items around the inflection points (support points) was estimated. This pattern was expected, on the basis of the sharply peaked hazard function (Figure 1). The estimates of support points and posterior class proportions are summarized in Table 2. As shown there, the support points were quite evenly distributed, and were close to the peaks of the empirical hazard function (Figure 1). As indicated by the estimated posterior proportions, the distribution of the steps variable appeared to have two modes. For $c = 6$, the support point was somewhat above the maximum test length of 20 items (i.e., $\hat{\delta}_6 = 20.92$) and this class received a proportion of $\hat{\pi}_6 = .19$. This class describes test takers who were likely to complete the test. Class $c = 3$ received the largest proportion, indicating that around 30% of the test takers ($\hat{\pi}_3 = .32$) switched to NRIs around the middle position of the test ($\hat{\delta}_3 = 10.43$).

Comparison With the Model of Glas and Pimentel (2008). We now turn to the comparison of the MDSEM, in which the indicator of group membership was discarded, with the model for NRIs proposed by Glas and Pimentel (2008). The first step was to estimate the MDSEM without considering group membership and to estimate Glas and Pimentel's model. The MDSEM contained 57 free parameters and achieved a log-likelihood value of LL = −6,994 (AIC = 14,103, BIC = 14,381). The model of Glas and Pimentel (2008) contained fewer parameters (44) and achieved a lower log-likelihood value of LL = −7,172 (AIC = 14,432, BIC = 14,647). In addition, the information indices were clearly in favor of the MDSEM.

Despite the difference in model-data fit, both models provided nearly identical estimates for the measurement part of the proficiency variable. The estimates of item discriminations showed only minor deviations between the models (mean absolute

Table 1. Fit Statistics for Mixture Discrete (Item) Sequence Event Model With Different Numbers of Latent Classes (K).

	No. of parameters	Log-likelihood	AIC	BIC
$K = 3$	51	−7,066	14,233	14,482
$K = 4$	55	−7,056	14,222	14,490
$K = 5$	59	−7,021	14,159	14,447
$K = 6$	63	−6,969	14,064	14,371
$K = 7$	67	−6,963	14,060	14,387
$K = 8$	71	−6,954	14,050	14,396

Note. AIC, Akaike information criterion; BIC, Bayesian information criterion.

Table 2. Proportions and Support Points of the Nonparametric Representation of the Distribution of the Steps Variable.

	$\hat{\pi}_k$	$\hat{\delta}_k$	(SE)
$k = 1$	.05	4.75	(0.167)
$k = 2$	.16	7.65	(0.078)
$k = 3$	.32	10.43	(0.052)
$k = 4$	.19	13.48	(0.064)
$k = 5$	.09	17.62	(0.099)
$k = 6$	.19	20.92	(0.123)

deviation [MAD] = 0.030), and the same was true for item difficulties (MAD = 0.080). Furthermore, the variance of the latent proficiency variable was estimated by the model of Glas and Pimentel (2008) as $\hat{\sigma}^2_\theta = 1.95$ ($SE = 0.257$), and by the MDSEM as $\hat{\sigma}^2_\theta = 1.91$ ($SE = 0.248$). Marked differences were found for the estimated relationship between θ and the steps variable. In the model of Glas and Pimentel (2008), the unstandardized regression weight was estimated to be $\hat{\gamma}_{\xi,\theta} = -0.21$ ($SE = 0.044$, $p < .001$) and a standardized counterpart to be $\hat{\gamma}^{stnd}_{\xi,\theta} = -0.39$ ($SE = 0.100$, $p < .001$). This result documents a relationship of medium strengths that indicates that higher levels of proficiency were related to earlier onsets of NRIs.

The multinomial logit coefficients determined by the MDSEM are reported in Table 3. The intercept parameters mirrored the latent class proportions (Table 2). The regression weights represent the change in the log-odds of belonging to class $c = l$ relative to the reference class $c = 6$ for one-unit increase in θ. The analyses uncovered a pattern reflecting a curvilinear relationship and indicating that the chance of being classified into Classes 2 to 4, as opposed to Class 6, increased with higher values of θ. To gain a better insight into the relationships, Table 3 also reports the class probabilities expected for values of θ at the 10th, 50th, and 90th percentiles of the (normal) proficiency distribution. As can be seen, test takers of low ability tended to be

Table 3. Multinomial Logit Coefficients of the Regression of Latent Class on the Proficiency Variable (θ), and Predicted Class Probabilities as Selected Values of θ (10th, 50th, and 90th Percentiles).

	Multinomial logit coefficients				Predicted class probabilities		
	$\hat{\nu}_k$	(*SE*)	$\hat{\gamma}_{k,\theta}$	(*SE*)	$\theta = -1.77$	$\theta = 0.00$	$\theta = 1.77$
$k = 1$	−1.25	(0.196)***	0.08	(0.184)	.08	.05	.03
$k = 2$	−0.05	(0.123)	0.38	(0.113)**	.15	.17	.17
$k = 3$	0.59	(0.107)***	0.49	(0.098)**	.23	.32	.40
$k = 4$	0.11	(0.117)	0.46	(0.104)**	.15	.20	.24
$k = 5$	−0.69	(0.148)***	0.20	(0.124)	.11	.09	.07
$k = 6$	—	—	—	—	.30	.18	.09

**$p < .01$. ***$p < .001$.

more evenly distributed across latent classes, whereas high ability test takers became more concentrated in the interim classes, especially in Class 3.

To facilitate a better comparison of the predictions made by the two models, Figure 2 provides the survival curves derived at the 10th, 50th, and 90th percentiles of the (normal) proficiency distribution. The survival functions were markedly different. More specifically, Glas and Pimentel's (2008) model predicted that the survival curves were already different at the onset of the first NRIs (i.e., starting from $j = 5$). In contrast, the MDSEM revealed that the onset of NRIs, and hence the survival curves, started to be affected by proficiency from the middle position (around $j = 10$) on, which means that the occurrence of NRIs in the second quarter of the test (between the 5th and 10th item position) was not related to proficiency. In addition, the MDSEM indicated larger differences in the survival probabilities in the last quarter of the test, compared to the model of Glas and Pimentel (2008). Moreover, the survival curves provided by the MDSEM were not as smooth as the curves provided by the model of Glas and Pimentel (2008), which were close to a linear function. The survival curves of the MDSEM appeared to reflect the peaked nature of the hazard functions (Figure 1).

Differential Onset of NRIs. The last issue considered here concerns the question of whether membership in the two VET fields was related to the onset of NRIs. Hence, we now turn to the full MDSEM, in which the group membership was included (0 = *technicians*, 1 = *clerks*). The goodness of fit of the full MDSEM is reported in Table 1. In order to find out whether group membership was related to the onset of NRIs over and above proficiency, we estimated a more constrained version of the MDSEM, in which the multinomial weights of group membership were fixed to zero. The fit of the models was compared via an LRT and provided a statistically significant result, $\chi^2(df = 5) = 24.9$ ($p < .001$), indicating that group membership was related to the onset of NRIs over and above the proficiency variable.

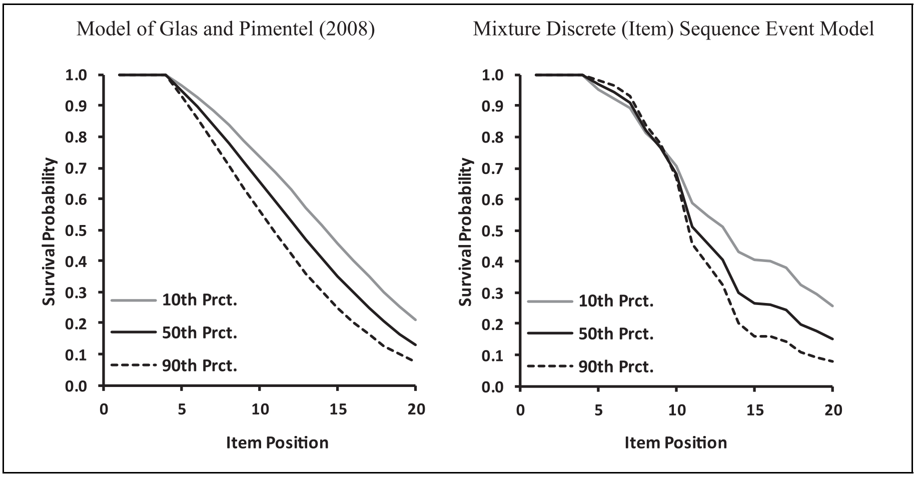

Figure 2. Survival functions determined for the 10th, 50th, and 90th percentile (Prct.) of the proficiency distribution determined on basis of the model of Glas and Pimentel (2008) and the Mixture Discrete (Item) Sequence Event Model (MDSEM).

The analysis provided the expected results. Clerks were found to have a higher proficiency level ($\hat{\gamma}_{\theta,x}$ = 0.66, *SE* = 0.130, $p < .001$), and a lower probability of belonging to the classes $c = 2$ and $c = 3$ than to the reference class $c = 6$, compared with the technicians. These results are shown by the multinomial regression weights provided in Table 4.

The regression weights for the proficiency variable predicting class membership were very similar to the results provided in Table 3. Table 4 also reports the class probabilities predicted by proficiency and group membership. The corresponding probabilities were evaluated at the 10th, 50th, and 90th percentiles of the combined proficiency distribution with equally weighted groups. Clerks of the same proficiency level were more likely to belong to the latent classes associated with a later onset of NRIs. This finding is visualized by the survival functions in Figure 3. The survival function already differed between groups right after the first onset of NRIs (at $j = 5$). In this region, survival did not depend on proficiency. The most pronounced group differences were determined for the third quarter of the test, where the survival function showed a steeper decrease at all levels of proficiency in the group of technicians. In the fourth quarter of the test, the survival curve was flatter for technicians, but the survival probability was still lower compared to the group of industrial clerks.

Summary

With this application, we intended to provide an example for an application of the proposed MDSEM in a low-stakes test characterized by a high prevalence of NRIs.

Table 4. Multinomial Logit Coefficients of the Regression of Latent Class on the Proficiency Variable (θ), and Predicted Class Probabilities (PCP) as Selected Values of θ (10th, 50th, and 90th Percentiles in the Combined Sample) for Subgroups.

	Multinomial logit coefficients						PCP: Technicians			PCP: Industrial clerks		
	$\hat{\nu}_k$	(*SE*)	$\hat{\gamma}_{k,\theta}$	(*SE*)	$\hat{\gamma}_{k,x}$	(*SE*)	$\theta = -1.45$	$\theta = 0.33$	$\theta = 2.12$	$\theta = -1.45$	$\theta = 0.33$	$\theta = 2.12$
$k = 1$	−1.10	(0.206)***	0.14	(0.190)	−0.80	(0.523)	.08	.05	.03	.04	.03	.02
$k = 2$	0.05	(0.138)	0.42	(0.118)**	−0.67	(0.305)*	.16	.18	.19	.10	.13	.15
$k = 3$	0.67	(0.122)***	0.54	(0.103)**	−0.74	(0.257)**	.25	.35	.45	.15	.24	.33
$k = 4$	0.10	(0.135)	0.48	(0.109)**	−0.21	(0.266)	.15	.20	.22	.16	.22	.28
$k = 5$	−0.88	(0.181)***	0.15	(0.130)	0.52	(0.315)	.09	.07	.04	.20	.16	.11
$k = 6$	—	—	—	—	—	—	.28	.15	.07	.35	.22	.11

*$p < .05$. **$p < .01$. ***$p < .001$.

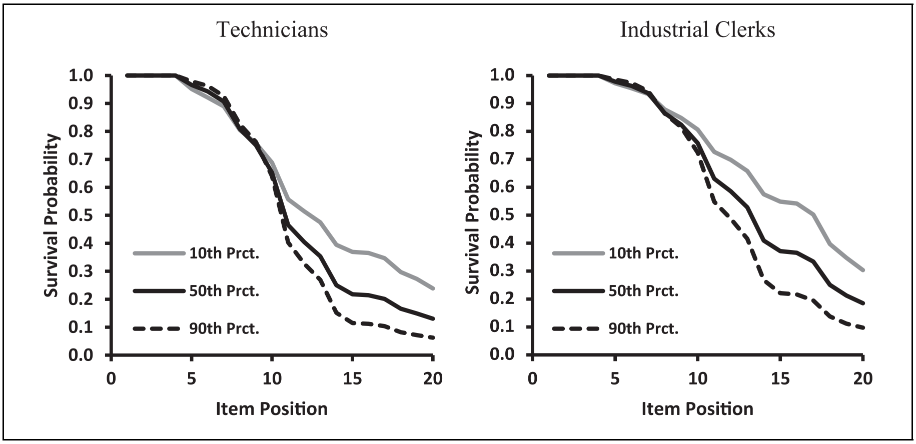

Figure 3. Survival functions determined for the 10th, 50th, and 90th percentiles (Prct.) of the joint proficiency distribution, determined for subgroups on the basis of the Mixture Discrete (Item) Sequence Event Model (MDSEM).

As we have shown, the MDSEM provides a method for detecting nonlinear patterns of the onset points of NRIs by using a semiparametric parameterization of the steps model. Compared with the parametric NRI model provided by Glas and Pimentel (2008), the MDSEM allows for a more flexible representation of the test survival function. In the present case, the MDSEM provided a survival curve that better reflected the peaked nature of the hazard function. In addition, the MDSEM does not rely on the proportional hazard assumption and was therefore able to identify regions where the onset of NRIs depended on person variables, and regions where NRIs did not depend on the variables considered. Finally, as we have demonstrated, the MDSEM allows for a simple test of the differential onset of NRIs as a function of the covariates, while simultaneously controlling for the proficiency variable.

Simulation Study

In this section, we report the results of a simulation study that was conducted in order to study the behavior of the MDSEM in the presence of a high amount of NRIs. We examined the MDSEM's capability of uncovering (1) item parameters (i.e., discriminations, α, and difficulties, β, Equation 2), (2) structural parameters pertaining to the relationship between a covariate x and the proficiency variable θ (i.e., $\gamma_{\theta,x}$, Equation 14) and the variance of θ conditional on x (i.e., $\sigma^2_{\zeta_\theta}$, Equation 14), and (3) the survival function $P(S_i \geq m|\theta_i, x_i)$ (Equation 16). We simulated item responses and response indicators for a test with $J = 30$ items administered to two groups (variable x) of equal size ($N = 1{,}000$ per group). Three conditions with 100 replications per condition were examined. In the first condition, NRIs depended solely on group membership. In the second condition, NRIs were only affected by the proficiency variable θ. Finally, in the third condition, NRIs depended on both characteristics. Except when constrained to be zero, the effects of x and θ on the onset of NRIs were held constant across conditions. The data sets were generated in such a way that all subjects had complete data on the first four items, and approximately 33% of the item responses were missing in all conditions.

The probabilities of NRI onsets were generated via a nonnormally distributed steps variable, δ (Equation 8) with λ fixed to one in the data-generation process. Nonnormality in δ was generated by means of a mixture of $K = 6$ univariate normal distributions with proportions $\pi_1 = .12$, $\pi_2 = .10$, $\pi_3 = .21$, $\pi_4 = .28$, $\pi_5 = .16$, and $\pi_6 = .14$, means $\mu_1 = 7$, $\mu_2 = 12$, $\mu_3 = 17$, $\mu_4 = 22$, $\mu_5 = 27$, and $\mu_6 = 35$, and variances $\sigma_1^2 = 1.00$, $\sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 2.25$, and $\sigma_6^2 = 12.25$. The relationships of δ with θ and x were generated via a multinomial logit model (Equation 15) with parameters $\gamma_{1,\theta}$ = -2.00, $\gamma_{2,\theta}$ = 0.50, $\gamma_{3,\theta}$ = 1.50, $\gamma_{4,\theta}$ = 2.00, and $\gamma_{5,\theta}$ = -0.25 for θ, and $\gamma_{1,x}$ = -2.00, $\gamma_{2,x}$ = -0.75, $\gamma_{3,x}$ = -0.75, $\gamma_{4,x}$ = -2.00, and $\gamma_{5,x}$ = 2.00 for x. Intercepts were set in such a way that the proportions π_1 to π_6 were equal across conditions.

The discrimination parameters for item responses α were generated according to a uniform distribution ranging between 0.5 and 1.5 (mean = 1), whereas the difficulty parameters β followed a standard normal distribution (mean = 0). We centered x to its mean and fixed the intercept of the structural regression part (Equation 14) to zero. The effect of x on θ was set to 0.5, and $\sigma_{\zeta_\theta}^2$ was set to 0.937. These values imply that θ had a variance of one and a mean of zero, so that the effect of x can be directly interpreted as a standardized effect.

In each condition, the data were analyzed via the MDSEM with $K = 6$ classes, and a 2PL model in which the response indicators were ignored. In both models, θ was regressed on x, and both models were identified by constraining the mean of the item discriminations to one. In light of the previous results, we expected the MDSEM to provide less biased parameters than a 2PL ignoring NRIs in conditions in which NRIs were nonignorable (i.e., Conditions 2 and 3). In Condition 1, we expected both models to perform equally well because the missing-data process was completely due to x. Because of its flexible nature, we expected the MDSEM to provide unbiased estimates of survival functions for various combinations of θ and x (10th, 50th, and 90th percentiles of the θ-distribution for each value of x).

Results

Table 5 provides the bias (i.e., the difference between average estimates and population values) of the parameters $\gamma_{\theta,x}$, and $\sigma_{\zeta_\theta}^2$, as well as the coverage rates of the parameter estimates (i.e., the proportion of parameter estimates whose 95% confidence intervals included the population values). The results for the 2PL model that ignored NRIs show that this model provided essentially unbiased structural parameters with good coverage rates only in the first condition. In this model, the variability of θ was underestimated in Conditions 2 and 3. Furthermore, the regression weight $\gamma_{\theta,x}$ was

Table 5. Population Values, Parameter Bias, and Coverage Rates for Structural Parameters Provided by the 2PL Ignoring NRIs, and the MDSEM Separated by Conditions (Condition 1: NRIs Affected by *x*; Condition 2: NRIs Affected by θ; Condition 3: NRIs Affected by *x* and θ).

		2PL		MDSEM	
	Population	Bias	Coverage	Bias	Coverage
Condition 1					
$\gamma_{\theta,x}$	0.500	0.001	1.00	0.001	1.00
$\sigma^2_{\zeta_\theta}$	0.937	0.011	.99	0.013	.99
Condition 2					
$\gamma_{\theta,x}$	0.500	−0.043	.98	0.005	1.00
$\sigma^2_{\zeta_\theta}$	0.937	−0.132	.18	0.028	.98
Condition 3					
$\gamma_{\theta,x}$	0.500	−0.093	.50	0.003	1.00
$\sigma^2_{\zeta_\theta}$	0.937	−0.120	.28	0.022	1.00

Note. $\gamma_{\theta,x}$ = regression weight of *x* predicting θ, $\sigma^2_{\zeta_\theta}$ = variance of θ conditional on *x*; 2PL = two-parameter logistic; MDSEM = Mixture Discrete (Item) Sequence Event Model; NRI = not-reached item.

clearly biased in the third condition. The bias in the estimate of $\gamma_{\theta,x}$ was lower in the second condition, where the model also provided an acceptable coverage rate. In contrast, the MDSEM provided virtually unbiased structural parameters that were accompanied by good coverage rates in all conditions studied.

Figure 4 provides scatter plots of the population values and the average item parameter estimates. Both models provided almost identical estimates that were virtually unbiased in the first condition. In the second and third conditions, where NRIs depended on θ, the MDSEM provided more accurate estimates for the items' discrimination parameters. In addition, the estimates of item difficulties appeared to be somewhat more accurate in the MDSEM, although the parameters provided by the standard 2PL were not strongly biased.

The last issue approached was the recovery of survival functions. The MDSEM correctly identified the variables not related to the onset of NRIs. Type I error rates for the multinomial logistic regression weights of θ were close to the nominal rate of .05 (range .06 to .02) in Condition 1, and the same pattern was found for the Type I errors of the logistic regression weights of x in Condition 2 (range .08 to .03). The survival functions provided by the model for the 10th, 50th, and 90th percentiles of the θ-distribution for both values of x are presented in Figure 5. As can be seen, the survival functions were virtually unbiased.

Summary

The results clearly show the advantages of the MDSEM. The model accurately estimated the survival functions in each condition studied, thereby underscoring the MDSEM's utility for examining the determinants of test takers' onset points of NRIs.

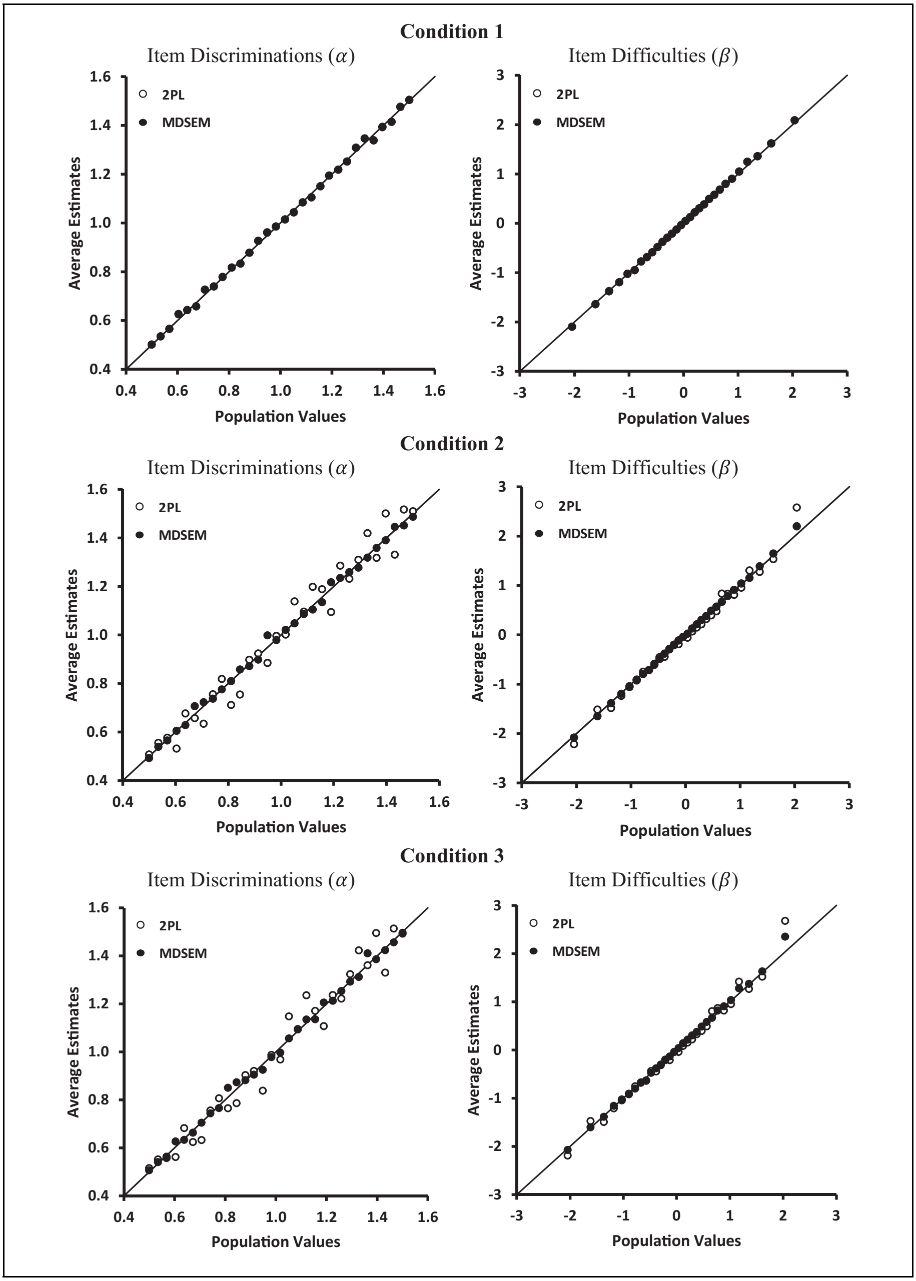

Figure 4. Estimated item parameters by corresponding population values for the 2PL ignoring NRIs and the MDSEM separated by conditions (Condition 1: NRIs affected by *x*; Condition 2: NRIs affected by θ; Condition 3: NRIs affected by *x* and θ).

Note. 2PL = two-parameter logistic; MDSEM = Mixture Discrete (Item) Sequence Event Model; NRI = not-reached item.

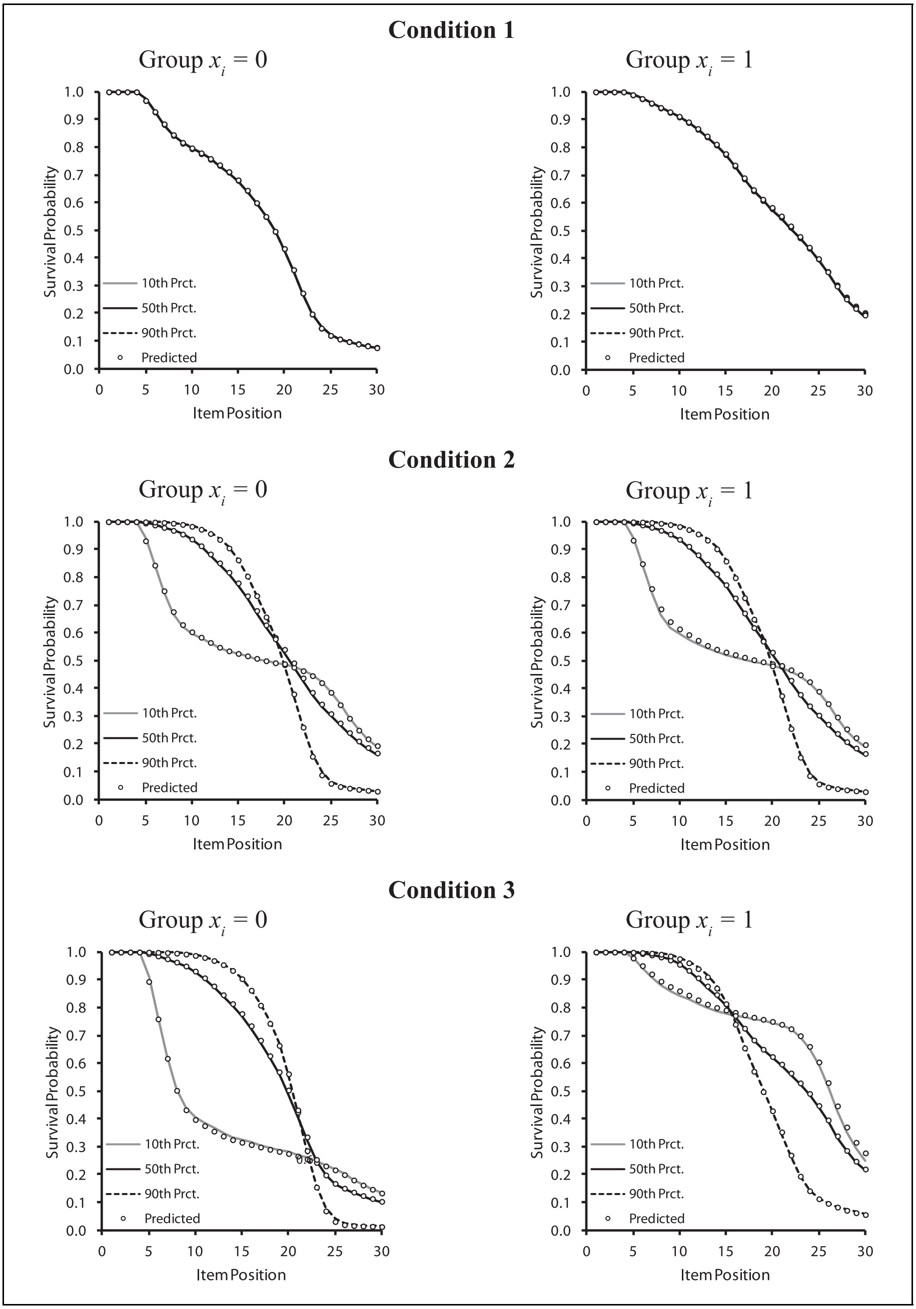

Figure 5. Population and estimated survival functions for the MDSEM separated by conditions (Condition 1: NRIs affected by *x*; Condition 2: NRIs affected by θ; Condition 3: NRIs affected by *x* and θ).

Note. MDSEM = Mixture Discrete (Item) Sequence Event Model; NRI = not-reached item.

In addition, our results show that the MDSEM reduced biases in parameters caused by nonignorable missing data. The MDSEM provided parameter estimates identical to the 2PL in a situation in which the missing-data process was accurately modeled by the inclusion of the covariate (i.e., Condition 1; Glas et al., 2015). In the conditions in which the onset of NRIs also depended on proficiency, the MDSEM provided unbiased estimates of the structural parameters, whereas the standard 2PL did not. In these conditions, the MDSEM produced more accurate item parameters, although the bias was also relatively low in the case of the conventional 2PL model that ignored NRIs.

Discussion

In educational assessments, one concern is whether the amount of NRIs is related to the proficiency being measured. Such relationships are considered to be indicative of NMAR patterns, which means that not accounting for such relationships could induce bias in the estimates of item parameters and students' proficiencies (Ludlow & O'Leary, 1999; for a review, see Pohl & Carstensen, 2013). However, an often-overlooked point is the possible relationship between NRIs and the student characteristics that are at the core of comparative studies. As we have argued in this article, situations in which such relationships cannot be accounted for by the students' proficiencies indicate a differential onset of NRIs that can be regarded as a threat to the validity of group comparisons. However, whether the differential onset of NRIs is treated as an indication of a threat to the validity of group comparisons, or whether it is treated as a key outcome in its own right, depends on the goals of the study.

Following this line of reasoning, we have presented the MDSEM as a flexible semiparametric approach that can be used for examining the differential onset of NRIs. Our model stands in close relationship with the approach suggested by Glas and Pimentel (2008) but relaxes some of its implicit assumptions, including the parametric distribution of the steps variable that assesses the onset point of NRIs, and the proportional hazards assumptions used for assessing the relationships of NRIs with the proficiency variable. The MDSEM proved valuable for determining the regions in which the NRIs were related to the explanatory variables, whereas this is not possible in the model proposed by Glas and Pimentel (2008).

The MDSEM has some similarities with the GDM suggested by Köhler et al. (2015b) for modeling the possibly nonnormal distribution of proficiency and the tendency to omit item responses. In contrast to the GDM, in the MDSEM, only the distributional assumptions for the steps variable are relaxed, while the proficiency variable is still assumed to be normally distributed. Furthermore, in our model, the categorization of δ is based on freely estimated support points, whereas these are defined in advance in the case of the GDM. The MDSEM is conceptually different because it allows NRIs to be predicted by proficiency and other covariates, whereas Köhler et al.'s (2015b) GDM allows only the relationship between proficiency and the tendency for omissions to be examined. In addition, the MDSEM has similarities with the model of Bacci and Bartolucci (2015), which uses freely estimated support points for proficiencies as well for the latent tendency to omit items. However, our model relies on a smaller number of parameters because we specified a variable that can be clearly interpreted as a steps variable, thereby making the interpretation of the model in real applications easier.

In summary, the MDSEM is easy to implement with conventional software packages, and it provided a better description of the datasets considered in this article than the model of Glas and Pimentel (2008). The MDSEM facilitates a straightforward test of the differential onset of NRIs by means of the LRT, and enables the presentation of these effects in a manner that can be easily understood by using the survival function borrowed from discrete-time event history analysis (Allison, 2014). As such, we believe that the method will prove useful in real applications concerned with the phenomenon of the differential onset of NRIs.

Furthermore, as we have shown in the simulation study, the MDSEM proved valuable for optimizing parameter estimates in the presence of NMAR patterns that were caused by NRIs. Compared with the 2PL model that ignored NRIs, the MDSEM clearly reduced biases in the variability of the proficiency variable and in group differences. As such, the MDSEM appears to be not only a valuable tool for examining whether NRIs are a threat to the validity of group comparisons, but also a model that helps to prevent such biases. However, this issue warrants further investigation. In particular, further studies should examine whether the MDSEM proves a viable alternative to existing models (e.g., Glas & Pimentel, 2008), as we think it does.

Future Developments

Although the MDSEM is highly flexible, it still includes assumptions, some of which can be easily relaxed. First, our hypothesis about the differential onset of NRIs was restricted to uniform effects, which means that the model assumed that respondents at all levels of proficiency were equally affected by this effect. Such a specification is common in other areas, for example, in studies investigating differential item functioning (DIF; Holland & Wainer, 1993). Following the DIF literature, the MDSEM could be extended to consider nonuniform effects by allowing the covariates to interact with the proficiency variable. Such models can be implemented in the case of categorical covariates by means of multigroup MDSEMs. Such an approach would make it possible to examine whether the effects of proficiency on the onset of NRIs differ between groups. In our opinion, evidence for an absence of the differential onset of NRIs would require group-invariant relationships between the proficiency variable and NRIs, as well as an absence of effects of the covariates on the onset of NRIs. Therefore, we decided to focus on uniform effects that can be interpreted more easily. However, extensions of the MDSEM that include interactions between the proficiency variable and covariates may be an interesting topic for further investigations.

A further restriction of the MDSEM is that it assumes the proficiency variable to be normally distributed and linearly related to the covariates. Given that these are standard assumptions in continuous latent variable models, we do not consider them to be a general shortcoming of the MDSEM. However, similar to the GDM proposed by Köhler et al. (2015b), in the case of omitted items, the distributional assumptions regarding the proficiency variable could be relaxed. The merits of relaxing the MDSEM should be clearly examined.

In addition, the MDSEM assumes an invariance of the measurement model applied to the item responses across groups and across patterns of NRIs. The first restriction can be easily relaxed in the context of multigroup models. We decided not to pursue this point, mainly for pragmatic reasons and to enhance the ease of presentation. Relaxing the invariance assumption across patterns of NRIs, however, requires other types of models, such as the pattern mixture models suggested in the context of the missing-data literature (Little, 1993). In this context, our proposed semiparametric approach for assessing the steps variable could be used to stratify the sample according to the relevant patterns of NRIs (see Rose et al., 2010). In a next step, by drawing on the stratified sample, the invariance assumption could be relaxed. Further studies could consider this issue.

Conclusion

NRIs reflect a type of test-taking behavior that could be of interest in substantive research. As we have argued in this article, NRIs can either be regarded as a key outcome variable or can perhaps be treated as a threat to the validity of group comparisons of proficiency levels. In this article, we present the MDSEM as a flexible and easy-to-use approach for studying the onset of NRIs. As we have demonstrated, the MDSEM can be analyzed using standard software, which might make this approach appealing for applied researchers.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Marit Kristine List http://orcid.org/0000-0001-6426-8143

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117-128.

Allison, P. D. (2014). *Event history and survival analysis*. Thousand Oaks, CA: Sage.

Bacci, S., & Bartolucci, F. (2015). A multidimensional finite mixture structural equation model for nonignorable missing responses to test items. *Structural Equation Modeling*, *22*, 352-365. doi:10.1080/10705511.2014.937376

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation* Modeling, *12*, 513-535. doi:10.1207/s15328007sem1204/1

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-549). Reading, MA: Addison-Wesley.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*, 309-319.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123-131.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*, 167-178. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol11/iss1/14

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, *38*, 87-111. doi:10.1111/j.2044-8317.1985.tb00818.x

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

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

Guo, J., Wall, M., & Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, *7*, 145–163. doi:10.1093/biostatistics/kxi046

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hutchison, D., & Yeshanew, T. (2009). Augmenting the use of the Rasch model under time constraints. *Quality & Quantity*, *43*, 717-772. doi:10.1007/s11135-007-9156-5

Kim, M., Vermunt, J., Bakk, Z., Jaki, T. F., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling*, *23*, 601-614. doi.org/10.1080/10705511.2016.1158655

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

Köhler, C., Pohl, S., & Carstensen, C. H. (2015b). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*, 850-874. doi: 10.1177/0013164414561785

Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance* (Research Report No. 93-49). Princeton, NJ: Educational Testing Service.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125-134. doi:10.2307/2290705

Little, R. J. A. (1994). A class of pattern-mixture models for multivariate incomplete data. *Biometrika*, *81*, 471-483.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21-39. doi:10.1037/1082-989X.10.1.21

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59*, 615-630. doi: 10.1177/0013164499594004

Masyn, K. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development*, *6*, 165-194.

Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, *30*, 27-58. doi:10.3102/10769986030001027

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535-569. doi:10.1080/10705510701575396

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling*, *23*, 782-797. doi:10.1080/10705511.2016.1221313

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

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Retelsdorf, J., Lindner, C., Nickolaus, R., Winther, E., & Köller, O. (2013). Forschungsdesiderate und Perspektiven—Ausblick auf ein Projekt zur Untersuchung mathematisch-naturwissenschaftlicher Kompetenzen in der beruflichen Erstausbildung (ManKobE) [Research desiderata and perspectives—Prospects for a project to study mathematical and scientific competencies in vocational education and training]. *Zeitschrift für Berufs- und Wirtschaftspädagogik, Beiheft*, *26*, 227-234.

Rose, N., von Davier, M., & Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychological Test and Assessment Modeling*, *57*, 472-498.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*, 795-819. doi:10.1007/s11336-016-9544-7

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

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37-50. doi:10.1016/j.intell.2014.10.003

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning for black examinees on Scholastic Aptitude Test analogy items* (Research Report No. 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1991). *Differential speededness and item omit patterns on the SAT* (Research Report No. 91-50). Princeton, NJ: Educational Testing Service.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, *18*, 155-195. doi: 10.2307/1165085

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English Language Learners. *Educational Assessment*, *13*, 108-131.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123-138). New York, NY: Springer.

Vermunt, J. K., & J. Magidson (2005). *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations.

von Davier, M. (2008), A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307. doi:10.1348/000711007X193957

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19-28.

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

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175-188. doi:10.2307/2531905