PSYCHOMETRIKA—VOL. 87, NO. 3, 835–867
SEPTEMBER 2022
https://doi.org/10.1007/s11336-021-09810-0

MODELING NOT-REACHED ITEMS IN TIMED TESTS: A RESPONSE TIME CENSORING APPROACH

JINXIN GUO AND XIN XU

BEIJING NORMAL UNIVERSITY

ZHILIANG YING

COLUMBIA UNIVERSITY

SUSU ZHANG

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Time limits are imposed on many computer-based assessments, and it is common to observe examinees who run out of time, resulting in missingness due to not-reached items. The present study proposes an approach to account for the missing mechanisms of not-reached items via response time censoring. The censoring mechanism is directly incorporated into the observed likelihood of item responses and response times. A marginal maximum likelihood estimator is proposed, and its asymptotic properties are established. The proposed method was evaluated and compared to several alternative approaches that ignore the censoring through simulation studies. An empirical study based on the PISA 2018 Science Test was further conducted.

Key words: not-reached items, missing data, response time, censoring, timed test.

In the ability testing literature, distinctions are often made between speed tests and power tests. In a pure speed test, it is assumed that all items can be responded correctly given enough time, and the amount of time it takes the participant to complete the items reflects ability. A pure power test, on the contrary, assumes that the items vary in difficulty and that the participants have an infinite amount of time to work on the items, and the response accuracy to the test items is used to measure ability. A large number of educational assessments, however, are a mixture of the two, where the items vary in difficulty, and a short enough time limit is imposed on the test so that examinees cannot take as much time as they desire (Cronbach and Warrington 1951; Roskam 1997). Time limits are imposed on many large-scale assessments (LSAs), such as the Trends in International Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP) in the United States, the Programme for International Student Assessment (PISA), and the National Educational Panel Study (NEPS) in Germany (e.g., OECD 2009; Pohl et al.2012; Rose et al.2010), as well as on high-stakes examinations, such as the Graduate Management Admission Test (GMAT), the Law School Admission Test (LSAT), and the United States Medical Licensure Exam (USMLE; e.g., Evans and Reilly 1972; Harik et al.2018). The imposition of a finite time limit can be attributed to both validity reasons, for example, to set the optimal speed level that examinees should operate under (Tijmstra and Bolsinova 2018), and financial considerations, for example, to ensure that enough tests can be administered in a given amount of time at a commercial testing center (Luecht and Sireci 2011).

For some timed tests, a nonnegligible proportion of students do not reach the end of the test by the time that it terminates, and responses to these not-reached items (NRIs) are thus missing.

Correspondence should be made to Susu Zhang, University of Illinois at Urbana-Champaign, Illinois, USA. Email: szhan105@illinois.edu

For example, on the NIO Intelligence Test in the Netherlands, Glas and Pimentel (2008) reported 27% of overall missing data due to NRI; in the 2006 PISA (OECD 2009; Pohl et al.2019; Rose et al. 2010), an average of 4% NRIs was observed across all participating countries, and for a specific country, the percent of missing due to NRIs ranged between .1% and 13%; for the German NEPS, Pohl et al. 2012 reported 13.46% of missing due to NRIs; for the GMAT, depending on the region, the percentage of individuals with at least one NRI in the verbal section ranged between 4.6% and 52.1% (Talento-Miller et al. 2013). While NRIs can also occur due to other reasons, such as early quitting (e.g., Ulitzsch et al. 2019), in the current article, we restrict our discussion to NRIs due to reaching the time limit.

In some operational tests (e.g., NAEP, Johnson and Allen 1992), missing responses due to NRIs are treated in the item calibration stage as if they were not administered to the participants, and they are treated as partially correct or incorrect in the scoring stage. Item parameters are estimated based on the observed responses only. This practice is supported by several simulation studies (Rose et al.2010; Pohl et al.2012), which suggested that IRT item parameter estimates are relatively robust against small-to-moderate amount of missing. An implicit assumption behind this practice is that the missing data due to NRIs can be treated as ignorable. However, as corroborated by extensive previous research, individuals' tendency to miss items due to the time limit is often related to their latent ability (e.g., Glas and Pimentel 2008; Holman and Glas 2005; Rose et al. 2010), and thus, the missingness is likely to be nonignorable.

Instead of treating the observed data as the complete data in parameter estimation, various model-based methods were also proposed to account for possibly nonignorable missingness in the presence of NRIs, where the missing propensity is incorporated into the measurement model. One approach is to use multidimensional IRT (MIRT) models to jointly account for the missing patterns and the observed responses (e.g., O'Muircheartaigh and Moustaki 1999; Moustaki and Knott 2000; Glas and Pimentel 2008; Holman and Glas 2005). Under this approach, the usual latent ability parameter governs the probability that a test taker gives a specific response, and a separate person-specific latent variable for missing propensity is introduced to model the probability that the test taker misses an item. The joint distribution of the latent ability and the missing propensity is also modeled. Parameter estimation under this approach takes both the missing patterns and the observed responses as input and jointly estimates the test takers' latent ability and latent missing propensity. Another common approach to modeling the missing patterns is via latent regressions (e.g., Rose et al. 2010; 2017). Instead of introducing a separate latent parameter for missing tendencies, functions of the missing data patterns, such as the percentage of NRIs or groupings based on NRI patterns, are used as a covariate for modeling the latent ability parameter. The conditional distribution of the latent ability, given the missing patterns, is thus taken into account in parameter estimation.

Most recently, response time-based approaches were proposed to model nonignorable missingness due to NRIs and other causes (Pohl et al. 2019; Lu et al.2018; Ulitzsch et al. 2019). The increased prevalence of computer-based assessments has enabled easy collection of reaction time data on each item, that is, the amount of time an examinee spends on a test item. Response times (RTs) provide an additional source of information on top of response accuracy and have been used for improving the estimation accuracy of item parameters and examinees' latent traits, understanding individuals' test-taking behavior and the test items' characteristics, and differentiating examinees using different test-taking strategies. To account for missingness due to NRIs, the hierarchical framework proposed by van der Linden (2007) is used to jointly model the observed responses and RTs on the reached items. Under this hierarchical framework, examinees' response accuracy to each item depends on a latent ability parameter through an IRT model, and the examinee's RT on each item follows a lognormal distribution (van der Linden 2006) that depends on a latent speed parameter. The joint distribution of the latent speed and the latent ability is further accounted for in the structural model. The rationale behind using a joint model for responses

and RTs to handle NRIs is as follows: Examinees' latent speeds affect how many items they can complete within the time limit and hence contain information about the amount of missingness due to NRIs. The joint modeling of responses and RTs takes into account the information on the missingness due to NRIs in speed estimation, and because speed and ability are often correlated, the missingness information is also incorporated into the estimation of latent ability and item parameters.

The availability of RT data opens up new possibilities in the modeling of missingness due to NRIs. One of the issues with modeling only the observed responses and RTs, however, is that it ignores the censoring due to the time limit. Suppose an individual's response to item j on a timed test is observed. Whether the response to item $j + 1$ is missing due to censoring clearly depends on whether the time that the individual would spend on item $j + 1$ exceeds the remaining time on the test. The missingness due to NRIs and the unobserved RTs thus remain to be dependent after conditioning on the observed responses and RTs. More specifically, in the presence of NRIs, the number of observed responses and RTs for each participant depends on the (random) number of reached items for the participant. Therefore, in addition to the observed responses and RTs on the reached items, the likelihood also depends on the number of reached items for the participant, which can be explained by the censoring of cumulative RTs. Ignoring this censoring mechanism in parameter estimation can lead to biased estimates of and incorrect inferences about the RT model item parameters and examinees' latent speeds, and, in the case that ability and speed are correlated, estimation and inferences of the latent ability and item response model parameters will also be affected. The present paper addresses this issue by introducing a framework under which the correct likelihood function can be written explicitly. Specifically, under a general class of joint models for responses and RTs, when the missingness is completely due to reaching the time limit, the explicit likelihood can be written without further assumptions on the missing mechanisms.

The rest of the paper is organized as follows: Section 1 introduces the proposed modeling framework for missing data due to NRIs using RT censoring. The likelihood of the observed data that accounts for the right censoring is given. A marginal maximum likelihood estimation (MMLE) method is subsequently proposed, and its asymptotic properties are established. Section 2 provides the details and results of a simulation study, which evaluates and compares parameter estimation using the proposed and other approaches. Results of an empirical study using the PISA 2018 Science data are subsequently presented in Sect. 3.

1. Modeling Missingness due to NRIs with RT Censoring

Consider a timed test administered to N examinees with a total of J items, where the items are presented to each examinee in a sequential order, that is, the next item is presented to the examinee when s/he submits the answer to the current question. This setup is similar to many computer-based timed tests (e.g., Glas and Pimentel 2008). To focus on the missing mechanisms of NRIs, let us further assume that the test does not allow other forms of missingness, such as omission and early-stopping. For each examinee i, suppose the test has a time limit $c_i < \infty$, and any items that the examinee has not responded to by time c_i are recorded as missing. Following Little and Rubin (1986), denote the complete response and response time data by $N \times J$ matrices $\mathbf{X} = (X_{ij})_{1\le i\le N,\ 1\le j\le J}$ and $\mathbf{T} = (T_{ij})_{1\le i\le N,\ 1\le j\le J}$, respectively, where $X_{ij} \in \{0, 1\}$ and $T_{ij} > 0$ are the response and response time, respectively, of the ith examinee to the jth item if there is no missingness. For examinee i, let θ_i and τ_i denote the latent ability and speed, which are assumed to be jointly normally distributed with the following specification,

$$\begin{pmatrix}\theta_i\\ \tau_i\end{pmatrix} \sim MVN\left(\mathbf{0}, \begin{pmatrix}1 & \rho\sigma_\tau\\ \rho\sigma_\tau & \sigma_\tau^2\end{pmatrix}\right). \tag{1}$$

For examinee i, assume that conditioning on latent traits θ_i and τ_i, the responses X_{ij} and response times T_{ij}, $j = 1, ..., J$ are mutually independent, i.e., the local independence holds, with the following distributional specifications

$$P(X_{ij} = 1 \mid \theta_i, \tau_i) = H(a_j\theta_i - b_j) = \frac{1}{1 + \exp[-(a_j\theta_i - b_j)]}, \tag{2}$$

where $H(z) = 1/(1 + e^{-z})$ and a_j and b_j are item parameters (Birnbaum 1968), and

$$\log T_{ij} \mid \theta_i, \tau_i \sim N\left(\gamma_j - \tau_i, \frac{1}{\alpha_j^2}\right), \tag{3}$$

where γ_j and α_j are the time-intensity and time-discrimination parameters, respectively (see van der Linden 2006; van der Linden 2007).

If examinee i had unlimited time to attempt the test, then X_{ij} and T_{ij} would have been observed for all j. However, given a finite total time limit c_i, the potential response and RT on any item s/he cannot respond to by time c_i will be censored and unobserved. In many scenarios, a universal time limit is imposed on a test, and c_i is the same for all examinees. There are, however, situations in which the amount of time permitted differs across participants. An example is when examinees randomly assigned to different test forms are allowed different time limits: Using precalibrated RT model item parameters, time limit may be accordingly set on a test to control the risk that a participant runs out of time (van der Linden 2011). As a result, time limits set in this manner can vary across test forms consisting of different items whose RT parameters differ. In the presence of finite time limits, the observed data for each i and j are the triplet $(R_{ij}, \tilde{X}_{ij}, \tilde{T}_{ij})$, where $R_{ij} = \mathcal{I}(\sum_{l=1}^{j} T_{il} \le c_i)$, $\tilde{X}_{ij} = X_{ij}R_{ij}$ and $\tilde{T}_{ij} = T_{ij} \wedge (c_i - \sum_{l=1}^{j-1} T_{il})^+$, where, for real numbers u and v, $u^+ = \max\{u, 0\}$ and $u \wedge v = \min\{u, v\}$. Note that $\tilde{X}_{ij} = X_{ij}$ and $\tilde{T}_{ij} = T_{ij}$ when $R_{ij} = 1$, that is, when the item is reached. We let $\boldsymbol{R}_i = (R_{ij})_{1\le j\le J}$ be the missing indicator vector of examinee i and $\boldsymbol{R} = (R_{ij})_{1\le i\le N, 1\le j\le J}$ be the missing indicator matrix of all examinees on all items, and similarly define $\tilde{\boldsymbol{X}}_{\boldsymbol{i}}$, $\tilde{\boldsymbol{T}}_{\boldsymbol{i}}$, $\tilde{\boldsymbol{X}}$, and $\tilde{\boldsymbol{T}}$.

Let $S_i = \sum_{j=1}^{J} R_{ij}$ be the number of items examinee i has completed by time c_i, and let $c_{i,j} = c_i - \sum_{l=1}^{j-1} T_{il}$ be the remaining time allowed for item j. Then the following proposition holds.

Proposition 1. *Given the ith test taker's latent traits (θ_i, τ_i), the joint density of $\boldsymbol{R}_i$, $\tilde{\boldsymbol{X}}_i$, and $\tilde{\boldsymbol{T}}_i$ is*

$$\prod_{j=1}^{S_i}\left[P(\tilde{X}_{ij} \mid \theta_i, \tau_i) f(\tilde{T}_{ij} \mid \theta_i, \tau_i)\right]\left[\bar{F}(c_{i,S_i+1} \mid \theta_i, \tau_i)\right]^{\mathcal{I}(S_i<J)},$$

where $P(\tilde{X}_{ij} \mid \theta_i, \tau_i)$ is the response probability based on Equation (2), and $f(\tilde{T}_{ij} \mid \theta_i, \tau_i)$ is the RT density based on Equation (3), with $\bar{F}(t \mid \theta_i, \tau_i) = \int_t^\infty f(s \mid \theta_i, \tau_i)\, ds$ as its survival function.

The above proposition is an adaptation of the standard results on the likelihood of right-censored survival data (see, Lawless 2011), and a proof is included in Appendix I. It should be noted that the current framework assumes the person-specific time limit ci to be fixed instead of random. The proposition hence does not generalize to cases where ci is random and correlated with individual latent traits. Define item parameter vectors $\boldsymbol{a} = (a_j)_{1\le j\le J}$, $\boldsymbol{b} = (b_j)_{1\le j\le J}$, $\boldsymbol{\alpha} = (\alpha_j)_{1\le j\le J}$ and $\boldsymbol{\gamma} = (\gamma_j)_{1\le j\le J}$.

From Proposition 1, by integrating out the latent trait variables, we get the following (marginal) likelihood function:

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\tau, \rho) = \prod_{i=1}^{N} \left\{ \iint \prod_{j=1}^{S_i} \left[P(\tilde{X}_{ij} \mid \theta_i, \tau_i) f(\tilde{T}_{ij} \mid \theta_i, \tau_i) \right] \left[\bar{F}(c_{i,S_i+1} \mid \theta_i, \tau_i) \right]^{\mathcal{I}(S_i<J)} g(\theta_i, \tau_i \mid \sigma_\tau^2, \rho) d\theta_i d\tau_i \right\}, \tag{4}$$

where $g(\theta_i, \tau_i \mid \sigma_\tau^2, \rho)$ is the joint density of θ and τ as in Eq. (1).

Remark 1. (Untimed Test) For the special case of no-time limit (i.e., $c_i = \infty$), $R_{ij} = \mathcal{I}(\sum_{j'=1}^{j} T_{ij'} \le \infty) = 1$ for all j. Consequently, $\tilde{\mathbf{X}}_i = \mathbf{X}_i$, $\tilde{\mathbf{T}}_i = \mathbf{T}_i$, and $S_i = J$. The joint density in Proposition 1 becomes equivalent to the complete data density, given by

$$P(\boldsymbol{X}_i, \boldsymbol{T}_i \mid \theta_i, \tau_i) = \prod_{j=1}^{J} \left[P(X_{ij} \mid \theta_i, \tau_i) f(T_{ij} \mid \theta_i, \tau_i) \right]. \tag{5}$$

The additional term involving $\bar{F}(\cdot)$ in Proposition 1 is due to the censoring on the first unreached item.

Aside from jointly modeling $(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{T}}_i, S_i)$, another common approach to parameter estimation in the presence of NRIs relies on the conditional distribution of observed responses/RTs given the number of reached items (e.g., Glas and Pimentel 2008), for instance, $P(\tilde{\boldsymbol{X}}_i \mid S_i)$. Below, we present two remarks on the marginal likelihood of RT- or response-model parameters when the missingness is due to right-censoring of response times. Although the two likelihoods are difficult to implement in practice for parameter estimation, it is shown that the likelihood of RT model parameters can be written based on the observed RTs and the missingness patterns, whereas the likelihood of response model parameters cannot be written based only on the observed responses and missingness patterns. For this reason, RT information, when available, is needed to correctly account for the missing mechanism due to NRIs.

Remark 2. (Marginal likelihood for RT parameters) Given the number of reached items, S_i, the marginal density of the observed RTs, $\tilde{\boldsymbol{T}}_i$, is

$$f(\tilde{\boldsymbol{T}}_i \mid S_i) = \frac{f(\tilde{\boldsymbol{T}}_i, S_i)}{P(S_i)}. \tag{6}$$

Here,

$$
\begin{aligned}
f(\tilde{\boldsymbol{T}}_i = \boldsymbol{t}_i, S_i = s) &= P(T_{ij} = t_{ij}, j = 1, \ldots, s, S_i = s) \\
&= P(T_{ij} = t_{ij}, j = 1, \ldots, s, T_{S_i+1} > c - \sum_{j=1}^{s} t_{ij}) \\
&= \iint P(T_{ij} = t_{ij}, j = 1, \ldots, s, T_{S_i+1} > c - \sum_{j=1}^{s} t_{ij} \mid \theta, \tau) g(\theta, \tau) d\theta d\tau \\
&= \int \prod_{j=1}^{s} f(t_{ij} \mid \tau) \bar{F}(c - \sum_{j=1}^{s} t_{ij} \mid \tau) g_\tau(\tau) d\tau, \text{ and} \\
P(S_i = s) &= \int_{t_1 + \ldots + t_s < c} f(t, s) dt, \text{ where the integration is over } \quad t_1 + \ldots + t_s < c, \\
&= \int_{t_1 + \ldots + t_s < c} \ldots \int \int \prod_{j=1}^{s} f(t_j \mid \tau) \bar{F}(c - \sum_{j=1}^{s} t_j \mid \tau) g_\tau(\tau) d\tau dt_1 \ldots dt_s,
\end{aligned}
$$

where $g_\tau(\tau) = g(\tau \mid \sigma_\tau^2)$ is the latent speed density assuming a normal distribution with mean 0 and variance σ_τ^2.

Consequently, the marginal likelihood of RT parameters is given by:

$$
L(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\tau) = \prod_{i=1}^{N} \left\{ \frac{\int \prod_{j=1}^{s_i} f(\tilde{t}_{ij} \mid \tau) \bar{F}(c - \sum_{j=1}^{s_i} \tilde{t}_{ij} \mid \tau) g_\tau(\tau) d\tau}{\int_{t_1 + \ldots + t_{s_i} < c} \ldots \int \int \prod_{j=1}^{s_i} f(t_j \mid \tau) \bar{F}(c - \sum_{j=1}^{s_i} t_j \mid \tau) g_\tau(\tau) d\tau dt_1 \ldots dt_s} \right\}, \tag{7}
$$

which does not involve the responses ($\tilde{\boldsymbol{X}}$) and the response model parameters.

Remark 3. (Marginal likelihood for response model parameters) Similar to Remark 2, one can also write the marginal density of $\tilde{\boldsymbol{X}}_i$ given S_i,

$$
\begin{aligned}
P(\tilde{\boldsymbol{X}}_i \mid S_i) &= \frac{P(\tilde{\boldsymbol{X}}_i, S_i)}{P(S_i)} \\
&= \frac{\iint P(\tilde{\boldsymbol{X}}_i, S_i \mid \theta, \tau) g(\theta, \tau) d\theta d\tau}{P(S_i)} \\
&= \frac{\iint \prod_{j=1}^{S_i} P(\tilde{X}_{ij} \mid \theta) P(S_i \mid \tau) g(\theta, \tau) d\theta d\tau}{\int_{t_1 + \ldots + t_{s_i} < c} \ldots \int \int \prod_{j=1}^{s_i} f(t_j \mid \tau) \bar{F}(c - \sum_{j=1}^{s_i} t_j \mid \tau) g_\tau(\tau) d\tau dt_1 \ldots dt_s}
\end{aligned} \tag{8}
$$

The numerator in Eq. (8) cannot be further reduced because θ and τ are inseparable inside the double integral. One exception is when θ and τ are independent, in which case $g(\theta, \tau) = g_\theta(\theta) g_\tau(\tau)$. In general, when θ and τ are dependent, $P(\tilde{\boldsymbol{X}}_i \mid S_i)$ still involves RT model parameters, and the marginal likelihood of response model parameters $(\boldsymbol{a}, \boldsymbol{b})$ given $\tilde{\boldsymbol{X}}_i \mid S_i$ cannot be explicitly written.

The current framework is established on the hierarchical model (van der Linden 2007), which assumes that the RTs to each item, $T_{it}, t = 1, \ldots, T$ are conditionally independent given (θ_i, τ_i). In other words, the latent speed (τ_i) is assumed stationary throughout the test, and the time spent on item j, T_{ij}, does not depend on the remaining time on the test, $c_i - \sum_{j'=1}^{j-1} T_{ij'}$. However, as is discussed in the remark below, the proposed approach for constructing the likelihood in the presence of RT censoring can be generalized to other joint models for responses and RTs.

Remark 4. (Extensions to other joint models for responses and RTs) Let $\boldsymbol{\omega}_i$ be the vector of latent trait parameters for person i, which can include latent speed, accuracy, or other characteristics of the individual, and let $\boldsymbol{\eta}$ denote the vector of model parameters. The RT censoring-induced missingness can be incorporated into the likelihood for following three classes of models:

Case 1 Given individual latent traits, $\boldsymbol{\omega}_i$, the responses and the RTs to each item are locally independent, and the responses and RTs are conditionally independent (e.g., van der Linden 2007; Bolsinova and Tijmstra 2018; Tijmstra and Bolsinova 2018), with

$$P(X_{ij}, T_{ij} \mid \boldsymbol{\omega}_i) = P(X_{ij} \mid \boldsymbol{\omega}_i) f(T_{ij} \mid \boldsymbol{\omega}_i). \tag{9}$$

In this case, the marginal likelihood of the model parameters, in the presence of NRIs due to reaching the time limit, takes the form of

$$L_1(\boldsymbol{\eta}) = \prod_{i=1}^{N} \left\{ \int \prod_{j=1}^{S_i} \left[P(\tilde{X}_{ij} \mid \boldsymbol{\omega}_i) f(\tilde{T}_{ij} \mid \boldsymbol{\omega}_i) \right] \left[\bar{F}(c_{i,S_i+1} \mid \boldsymbol{\omega}_i) \right]^{\mathcal{I}(S_i < J)} g(\boldsymbol{\omega}_i) d\boldsymbol{\omega}_i \right\}. \tag{10}$$

Case 2 Conditioning on the individual's latent traits, $\boldsymbol{\omega}_i$, the responses and RTs across items are locally independent, and the joint model of responses and RTs is parameterized so that the response distribution depends on the RT on the item (e.g., Wang and Hanson 2005; Bolsinova et al.2017), i.e.,

$$P(X_{ij}, T_{ij} \mid \boldsymbol{\omega}_i) = P(X_{ij} \mid T_{ij}, \boldsymbol{\omega}_i) f(T_{ij} \mid \boldsymbol{\omega}_i). \tag{11}$$

In this case, the marginal likelihood of the model parameters incorporating NRIs takes the form of

$$L_2(\boldsymbol{\eta}) = \prod_{i=1}^{N} \left\{ \int \prod_{j=1}^{S_i} \left[P(\tilde{X}_{ij} \mid \tilde{T}_{ij}, \boldsymbol{\omega}_i) f(\tilde{T}_{ij} \mid \boldsymbol{\omega}_i) \right] \left[\bar{F}(c_{i,S_i+1} \mid \boldsymbol{\omega}_i) \right]^{\mathcal{I}(S_i < J)} g(\boldsymbol{\omega}_i) d\boldsymbol{\omega}_i \right\}. \tag{12}$$

Case 3 Conditioning on latent traits, $\boldsymbol{\omega}_i$, the responses and RTs are locally independent across items, and the joint distribution of responses and RTs is parameterized so that the RT distribution of an item depends on the item response (e.g., van der Linden and Glas 2010), i.e.,

$$P(X_{ij}, T_{ij} \mid \boldsymbol{\omega}_i) = P(X_{ij} \mid \boldsymbol{\omega}_i) f(T_{ij} \mid X_{ij}, \boldsymbol{\omega}_i). \tag{13}$$

In this case, the marginal likelihood of $\boldsymbol{\eta}$ incorporating NRI takes the form of

$$
\begin{aligned}
L_3(\boldsymbol{\eta}) = \prod_{i=1}^{N} \Bigg\{ \int \prod_{j=1}^{S_i} \Big[P(\tilde{X}_{ij} \mid \boldsymbol{\omega}_i) f(\tilde{T}_{ij} \mid \tilde{X}_{ij}, \boldsymbol{\omega}_i) \Big] \\
\times \Big[\bar{F}(c_{i,S_i+1} \mid X_{ij} = 1, \boldsymbol{\omega}_i) P(X_{ij} = 1 \mid \boldsymbol{\omega}_i) \\
+ \bar{F}(c_{i,S_i+1} \mid X_{ij} = 0, \boldsymbol{\omega}_i) P(X_{ij} = 0 \mid \boldsymbol{\omega}_i) \Big]^{\mathcal{I}(S_i < J)} g(\boldsymbol{\omega}_i) d\boldsymbol{\omega}_i \Bigg\}. \quad (14)
\end{aligned}
$$

1.1. Parameter Estimation

Marginal maximum likelihood (MML) estimation is used for estimating the parameters of the proposed model. Denote the set of parameters by $\boldsymbol{\eta} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\tau, \rho)$. The MML estimates of $\boldsymbol{\eta}$ are obtained by maximizing the logarithm of marginal likelihood function in Eq. (4). As it involves integration over the latent variables, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, an efficient way to estimate the parameters is through an iterative expectation-maximization (EM) algorithm. The detailed computational procedures employed in the EM algorithm can be found in Appendix II. For each examinee, the expected a posteriori (EAP) estimates of θ_i and τ_i can further be computed. The details for the EAP approximations are also given in Appendix II. Parameters of the marginal model for RTs with censoring can be estimated similarly.

1.2. Asymptotic Properties

Applying the standard results on the consistency and asymptotic normality of marginal maximum likelihood estimators (see Lehmann and Romano 2006), the difference between the MML estimates and true values of the item parameters, scaled by the root of the inverse of the Fisher information matrix, converges to the multivariate standard normal distribution. More precisely, denoting the item parameters of all items by $\boldsymbol{\zeta} = (\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{\alpha}', \boldsymbol{\gamma}')'$, we have

$$
\boldsymbol{I}_N(\boldsymbol{\zeta})^{\frac{1}{2}}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbb{1}_{4J}),
$$

where $\boldsymbol{I}_N(\boldsymbol{\zeta})$ is the Fisher information matrix with true parameters $\boldsymbol{\zeta}$, and $\mathbb{1}_{4J}$ is the $4J \times 4J$ identity matrix. The Fisher information function $\boldsymbol{I}_N$ is presented in Appendix III. The Fisher information $\boldsymbol{I}_N(\boldsymbol{\zeta})$ can be consistently approximated by: $\boldsymbol{I}_N(\hat{\boldsymbol{\zeta}})$, and it follows that

$$
\boldsymbol{I}_N(\hat{\boldsymbol{\zeta}})^{\frac{1}{2}}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbb{1}_{4J}).
$$

The standard error estimate ($\hat{\text{SE}}$) of $\hat{\boldsymbol{\zeta}}$ is subsequently given by:

$$
\hat{\text{SE}}(\hat{\boldsymbol{\zeta}}) = \sqrt{\text{diag}(\boldsymbol{I}_N(\hat{\boldsymbol{\zeta}})^{-1})}. \quad (15)
$$

2. Simulation Studies

2.1. Simulation Design

Simulation studies were performed to evaluate the parameter recovery of the proposed MML estimator and to compare its performance to several alternative methods, when the missing data mechanism can be completely explained by censoring due to time limit. The parameter estimation accuracy using different approaches, under varying degrees of test length, sample size, missing rate, and latent speed and ability correlation, is examined. In the presence of missingness due to NRIs, instead of directly incorporating the censoring process into the joint model for responses and RTs, other approaches may also be adopted for parameter estimation. In addition to the proposed joint model with censoring and marginal RT model with censoring, two alternative approaches were also evaluated in the current study, as described below.

Joint model for responses and RTs ignoring censoring. To handle data with NRI-induced missingness, a recently proposed method is to jointly model the observed responses and RTs under van der Linden's (2007) hierarchical model (e.g., Pohl et al.2019; Ulitzsch et al. 2019; Lu et al.2018). Adopting the notations in the current paper, the marginal likelihood given the observed responses and RTs using this approach is given by:

$$L(\mathbf{a}, \mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \rho, \sigma_\tau^2) = \prod_{i=1}^{N} \left\{ \iint \prod_{j=1}^{S_i} \left[P(\tilde{X}_{ij} \mid \theta_i, \tau_i) f(\tilde{T}_{ij} \mid \theta_i, \tau_i) \right] g(\theta_i, \tau_i \mid \sigma_\tau^2, \rho) \quad d\theta_i \, d\tau_i \right\}. \tag{16}$$

Marginal model for responses ignoring censoring. An approach adopted in many operational tests is to calibrate the item parameters based on the observed response data only (e.g., Johnson and Allen 1992). With this approach, the marginal likelihood of the IRT model parameters given the observed responses is given by:

$$L(\mathbf{a}, \mathbf{b}) = \prod_{i=1}^{N} \left[\int \prod_{j=1}^{S_i} P(\tilde{X}_{ij} \mid \theta_i) g_\theta(\theta_i) \, d\theta_i \right], \tag{17}$$

where g_θ is the probability density of the latent ability assuming a standard normal distribution.

This amounts to three total parameter estimation methods considered in the current study:

- Method 1, the proposed joint model of responses and RTs incorporating time censoring, which optimizes the logarithm of the likelihood in Eq. (4).
- Method 2, the joint model of observed responses and RTs ignoring the censoring, which optimizes the logarithm of the likelihood in Eq. (16).
- Method 3, the marginal model of observed responses only, without accounting for time censoring, which optimizes the logarithm of the likelihood in Eq. (17).

Performance of the three approaches was evaluated under different sample size, test length, latent trait distribution, and missingness conditions, specifically:

- Sample sizes of $N = 1000$, 4000, and 10000 were considered;
- Three conditions for the latent trait correlation were considered, namely ρ =-.4, 0, .8;
- Test lengths of $J = 10$ and 40 were considered;
- Three missingness conditions were considered: Under conditions 1 and 2 (C1 and C2), the proportion of examinees with at least one NRI was set as $r = 50\%$. The true RT model time intensity (γ_js) and variance ($1/\alpha_j$s) parameters were set higher in C1 than in C2, so that C1 will have higher proportion of NRIs in earlier items. Finally, under condition 3

TABLE 1.
True item parameters and response rates to each item under conditions C1–C3 for $J = 10$. **a** and **b** were set to the same values under the three conditions. $\boldsymbol{\alpha}_1$ and $\boldsymbol{\gamma}_1$ are the true parameters under C1 and C3, and $\boldsymbol{\alpha}_2$ and $\boldsymbol{\gamma}_2$ are the true parameters under C2

Item	Item parameters						Response rate		
	a	**b**	$\boldsymbol{\alpha}_1$	$\boldsymbol{\gamma}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\gamma}_2$	C1	C2	C3
1	0.93	1.60	1.21	−0.91	1.96	−0.94	1.000	1.000	1.000
2	1.68	0.77	1.23	−0.19	1.90	−0.24	1.000	1.000	1.000
3	1.11	−1.40	0.79	1.59	1.69	−0.90	0.911	1.000	0.989
4	1.82	−1.25	0.59	0.17	1.80	−0.66	0.858	1.000	0.981
5	1.91	−0.85	0.61	−0.29	1.02	−0.57	0.824	0.999	0.976
6	0.57	0.70	0.81	0.81	1.22	−0.62	0.758	0.997	0.970
7	1.29	0.46	0.55	1.29	1.64	1.45	0.592	0.937	0.917
8	1.84	0.74	0.65	−0.24	0.82	0.88	0.556	0.810	0.910
9	1.33	0.15	0.53	−1.47	0.98	−0.18	0.538	0.770	0.906
10	1.18	−0.66	0.89	0.26	0.85	1.79	0.500	0.500	0.900
Mean							0.754	0.901	0.955

(C3), the proportion of examinees with at least one NRI was 10%, and the RT model item parameters under C3 are the same as those in C1.

Under each simulation condition, the same set of true item and structural parameters were used. The true structural parameters for latent ability and speed distribution were set as: $\Sigma_{\theta\tau} = \begin{pmatrix} 1 & .5\rho \\ .5\rho & .25 \end{pmatrix}$, with $\rho \in \{-.4, 0, .8\}$ varying depending on the specific simulation condition. In other words, the standard deviation of latent speed, σ_τ, was set to .5.

500 sets of observed response and RT data for a time test were randomly generated. For each replication, N subjects' true ability and speed parameters were randomly drawn from MVN($\mathbf{0}$, $\Sigma_{\theta\tau}$). To generate the response and RT data with NRI, complete responses and RTs to the J items were generated based on the 2PL (Eq. (2)) and lognormal (Eq. (3)) models, respectively. The proportion of individuals with NRI (r) was controlled by choosing an appropriate time limit: For instance, under the $r = 50\%$ condition, where half of the examinees had at least one NRI, a universal time limit (c) was set for all participants as the 50th percentile of examinees' total RTs. Then, if an examinee's cumulative RT up to an item exceeds c, the response and RT for the item were masked and recorded as missing. In this way, on each simulated data set, approximately half or 10 percent of the examinees had at least one NRI. Table 1 presents the true item parameters and the response rates to each item under the $J = 10$ condition.

The EM algorithm was terminated when it reached the maximal number of iterations or when the change in log-likelihood fell below a fixed tolerance level. The estimates of $\boldsymbol{\eta}$ from the last iteration were retained as the final parameter estimates. For the fixed model parameters, the average bias, root-mean-squared error (RMSE), standard error (SE), average of standard error estimates (SEE), and coverage probability (CP) of the 95% confidence interval of the true value were computed across the 500 data sets. Specifically, for a parameter η, denote its estimate from the rth replication by $\hat{\eta}^{(r)}$, the following evaluation indices were computed:

$$\text{Bias}(\hat{\eta}) = \frac{1}{500}\sum_{r=1}^{500}(\hat{\eta}^{(r)} - \eta),$$

$$\text{RMSE}(\hat{\eta}) = \sqrt{\frac{1}{500}\sum_{r=1}^{500}(\hat{\eta}^{(r)} - \eta)^2},$$

$$\text{SE}(\hat{\eta}) = \sqrt{\frac{1}{500}\sum_{r=1}^{500}(\hat{\eta}^{(r)} - \bar{\hat{\eta}})^2},$$

$$\text{SEE}(\hat{\eta}) = \frac{1}{500}\sum_{r=1}^{500}\left(\widehat{\text{SE}}(\hat{\eta}^{(r)})\right), \text{ and}$$

$$\text{CP}(\hat{\eta}) = \frac{1}{500}\sum_{r=1}^{500}\mathcal{I}\left(\hat{\eta}^{(r)} - z_{.975}\widehat{\text{SE}}(\hat{\eta}^{(r)}) \leq \eta \leq \hat{\eta}^{(r)} + z_{.975}\widehat{\text{SE}}(\hat{\eta}^{(r)})\right),$$

where $\bar{\hat{\eta}} = \frac{1}{500}\sum_{r=1}^{500}\hat{\eta}^{(r)}$, $\widehat{\text{SE}}(\hat{\eta}^{(r)})$ is the standard error estimate from the rth replication computed based on Equation (15), and $z_{.975}$ is the .975 quantile of the standard normal distribution. Note that $\text{SEE}(\hat{\eta})$ is the mean of the estimated standard error based on the asymptotic properties of the estimator in Section 2.2, while $\text{SE}(\hat{\eta})$ is the observed standard deviation of $\hat{\eta}$ across the 500 replications. When the SEE provides a good approximation to the standard error of $\hat{\eta}$, $\text{SE}(\hat{\eta})$ and $\text{SEE}(\hat{\eta})$ are expected to be close. In addition to the fixed model parameters, the agreements between the examinees' true and estimated abilities and speeds were further examined. Because a different set of examinee latent traits were randomly generated in each replication, the MSE of the θ and τ estimates was computed in each replication separately, that is,

$$\text{MSE}(\hat{\theta}^{(r)}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i^{(r)} - \theta_i^{(r)})^2$$

and similarly for τ. The RMSE across replications is then reported, where

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{500}\sum_{r=1}^{500}\text{MSE}(\hat{\theta}^{(r)})},$$

and similarly for τ. To additionally examine how the latent trait estimates produced from the three methods differ in the ranking of test takers, for each pair of methods (e.g., M1 & M2, M1 & M3, and M2 & M3), the Kendall's rank correlation (denoted COR) between the latent trait estimates produced using the two methods was computed.

2.2. Results

2.2.1. Response model parameters The bias, RMSE, and SE of the item slope parameters (a_js), averaged across the last four items in each condition, are reported in Table 2. The overall difference among the three methods was small. However, a tendency for methods 2 and 3 (method 3 especially) to produce smaller estimates of item slope parameters was observed: For example, across all $N = 10000$ conditions, the bias with the proposed method, method 1, remained below .01. On the other hand, methods 2 and 3, which ignored the censoring term, displayed a consistent

TABLE 2.
Recovery of a parameters, averaged across the last four items, under different conditions using methods 1–3 (abbreviated M1–M3).

Condition	ρ	J	N	Bias			RMSE			SE		
				M1	M2	M3	M1	M2	M3	M1	M2	M3
C1	−0.4	10	1000	0.0207	0.0135	−0.0019	0.1973	0.1957	0.1937	0.1961	0.1951	0.1936
C1	−0.4	10	4000	0.0024	−0.0046	−0.0194	0.0943	0.0939	0.0950	0.0943	0.0937	0.0929
C1	−0.4	10	10000	−0.0005	−0.0073	−0.0221	0.0595	0.0597	0.0631	0.0596	0.0593	0.0590
C1	−0.4	40	1000	0.0201	0.0197	0.0004	0.1659	0.1658	0.1629	0.1648	0.1648	0.1629
C1	−0.4	40	4000	0.0030	0.0027	−0.0166	0.0821	0.0821	0.0828	0.0821	0.0821	0.0811
C1	−0.4	40	10000	0.0004	0.0000	−0.0192	0.0500	0.0500	0.0529	0.0500	0.0500	0.0493
C1	0	10	1000	0.0195	0.0194	0.0192	0.1966	0.1966	0.1965	0.1956	0.1956	0.1955
C1	0	10	4000	0.0027	0.0027	0.0026	0.0933	0.0933	0.0933	0.0933	0.0933	0.0933
C1	0	10	10000	−0.0001	−0.0001	−0.0001	0.0580	0.0580	0.0580	0.0580	0.0580	0.0580
C1	0	40	1000	0.0203	0.0203	0.0202	0.1675	0.1675	0.1674	0.1664	0.1664	0.1664
C1	0	40	4000	0.0043	0.0043	0.0043	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801
C1	0	40	10000	0.0015	0.0015	0.0015	0.0502	0.0502	0.0502	0.0502	0.0502	0.0502
C1	0.8	10	1000	0.0180	−0.0146	−0.0521	0.2025	0.1973	0.2059	0.2018	0.1969	0.1992
C1	0.8	10	4000	0.0012	−0.0304	−0.0693	0.0952	0.0979	0.1164	0.0953	0.0931	0.0934
C1	0.8	10	10000	−0.0005	−0.0322	−0.0709	0.0599	0.0669	0.0923	0.0600	0.0586	0.0589
C1	0.8	40	1000	0.0132	0.0107	−0.0556	0.1777	0.1772	0.1790	0.1773	0.1770	0.1702
C1	0.8	40	4000	0.0029	0.0003	−0.0652	0.0899	0.0897	0.1081	0.0899	0.0897	0.0863
C1	0.8	40	10000	0.0013	−0.0012	−0.0663	0.0557	0.0556	0.0852	0.0556	0.0555	0.0534
C2	−0.4	10	1000	0.0171	0.0151	−0.0079	0.1677	0.1673	0.1648	0.1669	0.1666	0.1647
C2	−0.4	10	4000	0.0003	−0.0017	−0.0242	0.0819	0.0819	0.0850	0.0818	0.0817	0.0808
C2	−0.4	10	10000	−0.0012	−0.0032	−0.0254	0.0517	0.0517	0.0574	0.0517	0.0516	0.0511
C2	−0.4	40	1000	0.0180	0.0175	−0.0051	0.1553	0.1552	0.1522	0.1544	0.1543	0.1522
C2	−0.4	40	4000	0.0041	0.0036	−0.0187	0.0757	0.0756	0.0768	0.0756	0.0756	0.0745
C2	−0.4	40	10000	0.0005	0.0000	−0.0223	0.0470	0.0470	0.0514	0.0470	0.0470	0.0462
C2	0	10	1000	0.0158	0.0159	0.0156	0.1668	0.1669	0.1668	0.1661	0.1661	0.1661
C2	0	10	4000	0.0014	0.0014	0.0013	0.0791	0.0791	0.0791	0.0790	0.0790	0.0791
C2	0	10	10000	−0.0006	−0.0006	−0.0006	0.0506	0.0506	0.0506	0.0507	0.0507	0.0507
C2	0	40	1000	0.0189	0.0189	0.0187	0.1555	0.1555	0.1555	0.1545	0.1545	0.1545
C2	0	40	4000	0.0044	0.0044	0.0044	0.0739	0.0739	0.0739	0.0738	0.0738	0.0738
C2	0	40	10000	0.0002	0.0002	0.0001	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467
C2	0.8	10	1000	0.0125	0.0044	−0.0605	0.1655	0.1644	0.1767	0.1652	0.1645	0.1654
C2	0.8	10	4000	−0.0010	−0.0090	−0.0762	0.0805	0.0806	0.1121	0.0806	0.0801	0.0811
C2	0.8	10	10000	−0.0011	−0.0091	−0.0766	0.0503	0.0508	0.0922	0.0503	0.0500	0.0502
C2	0.8	40	1000	0.0146	0.0113	−0.0687	0.1706	0.1699	0.1746	0.1700	0.1696	0.1605
C2	0.8	40	4000	0.0024	−0.0008	−0.0804	0.0838	0.0835	0.1130	0.0838	0.0836	0.0793
C2	0.8	40	10000	0.0000	−0.0032	−0.0821	0.0528	0.0528	0.0962	0.0527	0.0526	0.0499
C3	−0.4	10	1000	0.0149	0.0121	0.0075	0.1497	0.1492	0.1494	0.1490	0.1487	0.1493
C3	−0.4	10	4000	−0.0015	−0.0042	−0.0085	0.0714	0.0713	0.0719	0.0713	0.0712	0.0713
C3	−0.4	10	10000	−0.0015	−0.0042	−0.0084	0.0451	0.0451	0.0460	0.0451	0.0450	0.0452
C3	−0.4	40	1000	0.0098	0.0097	0.0048	0.1243	0.1243	0.1239	0.1241	0.1241	0.1239
C3	−0.4	40	4000	0.0027	0.0025	−0.0025	0.0611	0.0611	0.0608	0.0610	0.0610	0.0608
C3	−0.4	40	10000	−0.0001	−0.0002	−0.0054	0.0376	0.0376	0.0379	0.0376	0.0376	0.0375
C3	0	10	1000	0.0152	0.0152	0.0150	0.1509	0.1509	0.1508	0.1502	0.1502	0.1501
C3	0	10	4000	0.0003	0.0003	0.0003	0.0706	0.0706	0.0706	0.0705	0.0705	0.0705
C3	0	10	10000	−0.0014	−0.0014	−0.0015	0.0453	0.0453	0.0453	0.0453	0.0453	0.0453

TABLE 2.
continued

Condition	ρ	J	N	Bias			RMSE			SE		
				M1	M2	M3	M1	M2	M3	M1	M2	M3
C3	0	40	1000	0.0117	0.0117	0.0117	0.1233	0.1233	0.1233	0.1228	0.1228	0.1228
C3	0	40	4000	0.0029	0.0029	0.0029	0.0602	0.0602	0.0602	0.0601	0.0601	0.0601
C3	0	40	10000	0.0000	0.0000	0.0000	0.0376	0.0376	0.0376	0.0376	0.0376	0.0376
C3	0.8	10	1000	0.0113	0.0000	−0.0081	0.1441	0.1425	0.1470	0.1438	0.1425	0.1467
C3	0.8	10	4000	0.0008	−0.0104	−0.0208	0.0702	0.0704	0.0746	0.0702	0.0696	0.0716
C3	0.8	10	10000	−0.0006	−0.0118	−0.0222	0.0437	0.0449	0.0499	0.0438	0.0434	0.0446
C3	0.8	40	1000	0.0111	0.0103	−0.0047	0.1243	0.1241	0.1238	0.1238	0.1237	0.1238
C3	0.8	40	4000	0.0025	0.0018	−0.0131	0.0614	0.0614	0.0628	0.0614	0.0614	0.0614
C3	0.8	40	10000	0.0000	−0.0008	−0.0154	0.0387	0.0387	0.0417	0.0387	0.0387	0.0388

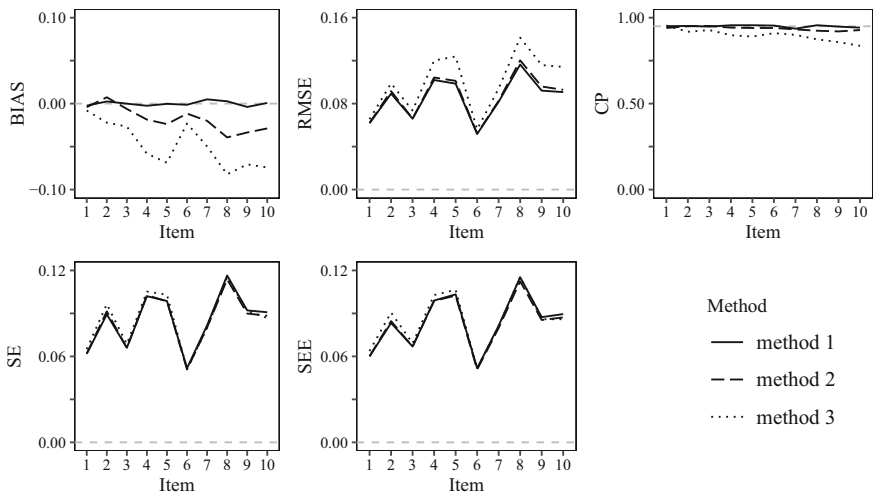

FIGURE 1.
Recovery of item discrimination parameters (i.e., a_j) under C1, $J = 10$, $N = 4000$, $\rho = .8$ in terms of bias, RMSE, and 95% CP, as well as the observed (SE) and estimated (SEE) standard errors, obtained from different estimation approaches.

negative average bias, especially when the proportion of missing was high (i.e., C1 and C2) and the absolute correlation between θ and τ was nonzero ($\rho = -.4$ and .8).

Figure 1 further presents the bias, RMSE, SE, SEE, and 95% CP of a_j estimates for each item, under condition C1 with $J = 10$, $\rho = .8$, $N = 4000$. Different line types are used to depict the results obtained using different methods. On each subfigure, the $x-$axis represents the indices of the 10 items, ordered according to their serial positions in the test. Items that appeared later were associated with lower completion rates. The gray, dashed lines in each of the subfigures provide a reference for the comparison of different methods, representing a value of 0 for bias, MSE, SE, and SEE and a value of 95% for CP. Method 1 produced accurate a_j estimates for all 10 items. Methods 2 and 3, however, displayed a tendency to underestimate the a_js, which is consistent with the observations from Table 2. The magnitude of the bias increased as the completion rate

TABLE 3.
Recovery of b parameters (for the last four items) under different conditions using methods 1–3 (abbreviated M1–M3).

Condition	ρ	J	N	Bias			RMSE			SE		
				M1	M2	M3	M1	M2	M3	M1	M2	M3
C1	−0.4	10	1000	0.0073	0.0418	0.1036	0.1204	0.1269	0.1582	0.1201	0.1195	0.1193
C1	−0.4	10	4000	0.0033	0.0373	0.0981	0.0596	0.0703	0.1148	0.0594	0.0593	0.0594
C1	−0.4	10	10000	−0.0003	0.0337	0.0944	0.0377	0.0505	0.1016	0.0376	0.0375	0.0375
C1	−0.4	40	1000	0.0015	0.0021	0.0455	0.1209	0.1208	0.1288	0.1208	0.1207	0.1202
C1	−0.4	40	4000	0.0002	0.0009	0.0438	0.0598	0.0598	0.0739	0.0598	0.0598	0.0595
C1	−0.4	40	10000	0.0008	0.0014	0.0443	0.0371	0.0371	0.0577	0.0371	0.0371	0.0368
C1	0	10	1000	0.0086	0.0082	0.0079	0.1216	0.1212	0.1208	0.1212	0.1208	0.1204
C1	0	10	4000	0.0025	0.0025	0.0023	0.0587	0.0587	0.0588	0.0586	0.0586	0.0587
C1	0	10	10000	0.0004	0.0004	0.0003	0.0380	0.0379	0.0378	0.0379	0.0378	0.0378
C1	0	40	1000	0.0027	0.0027	0.0026	0.1194	0.1194	0.1194	0.1193	0.1193	0.1193
C1	0	40	4000	0.0004	0.0004	0.0004	0.0599	0.0599	0.0599	0.0599	0.0599	0.0599
C1	0	40	10000	0.0004	0.0004	0.0004	0.0362	0.0362	0.0361	0.0362	0.0362	0.0362
C1	0.8	10	1000	0.0093	−0.0750	−0.1902	0.1256	0.1417	0.2237	0.1252	0.1203	0.1177
C1	0.8	10	4000	0.0031	−0.0800	−0.1951	0.0620	0.0999	0.2038	0.0619	0.0598	0.0587
C1	0.8	10	10000	−0.0007	−0.0838	−0.1983	0.0398	0.0922	0.2019	0.0398	0.0384	0.0375
C1	0.8	40	1000	0.0038	0.0005	−0.0950	0.1355	0.1350	0.1610	0.1353	0.1349	0.1299
C1	0.8	40	4000	0.0000	−0.0032	−0.0983	0.0666	0.0664	0.1170	0.0666	0.0663	0.0635
C1	0.8	40	10000	0.0007	−0.0026	−0.0982	0.0410	0.0409	0.1057	0.0410	0.0408	0.0391
C2	−0.4	10	1000	0.0043	0.0087	0.0563	0.1079	0.1081	0.1216	0.1078	0.1076	0.1067
C2	−0.4	10	4000	0.0019	0.0062	0.0530	0.0525	0.0528	0.0749	0.0524	0.0524	0.0519
C2	−0.4	10	10000	−0.0007	0.0036	0.0503	0.0334	0.0336	0.0609	0.0333	0.0333	0.0330
C2	−0.4	40	1000	0.0014	0.0022	0.0468	0.1146	0.1144	0.1228	0.1144	0.1142	0.1133
C2	−0.4	40	4000	0.0020	0.0029	0.0472	0.0564	0.0563	0.0731	0.0564	0.0563	0.0558
C2	−0.4	40	10000	0.0005	0.0012	0.0454	0.0353	0.0352	0.0573	0.0353	0.0352	0.0348
C2	0	10	1000	0.0050	0.0050	0.0050	0.1079	0.1079	0.1075	0.1077	0.1077	0.1074
C2	0	10	4000	0.0022	0.0022	0.0021	0.0519	0.0519	0.0519	0.0518	0.0518	0.0519
C2	0	10	10000	−0.0002	−0.0002	−0.0002	0.0331	0.0331	0.0330	0.0330	0.0330	0.0329
C2	0	40	1000	0.0013	0.0013	0.0013	0.1123	0.1123	0.1123	0.1122	0.1122	0.1122
C2	0	40	4000	0.0015	0.0015	0.0015	0.0563	0.0563	0.0563	0.0563	0.0563	0.0563
C2	0	40	10000	0.0002	0.0002	0.0002	0.0338	0.0338	0.0337	0.0338	0.0338	0.0337
C2	0.8	10	1000	0.0057	−0.0078	−0.0976	0.1072	0.1068	0.1468	0.1071	0.1065	0.1056
C2	0.8	10	4000	0.0012	−0.0121	−0.1021	0.0517	0.0530	0.1159	0.0517	0.0515	0.0512
C2	0.8	10	10000	−0.0004	−0.0137	−0.1035	0.0329	0.0356	0.1091	0.0329	0.0326	0.0322
C2	0.8	40	1000	0.0030	−0.0008	−0.0987	0.1288	0.1282	0.1561	0.1287	0.1281	0.1209
C2	0.8	40	4000	0.0011	−0.0023	−0.1009	0.0625	0.0622	0.1167	0.0625	0.0622	0.0587
C2	0.8	40	10000	−0.0002	−0.0038	−0.1022	0.0384	0.0383	0.1085	0.0384	0.0381	0.0362
C3	−0.4	10	1000	0.0055	0.0129	0.0247	0.0957	0.0964	0.0988	0.0955	0.0955	0.0956
C3	−0.4	10	4000	0.0029	0.0101	0.0215	0.0471	0.0481	0.0518	0.0470	0.0470	0.0471

of the item decreased. For all three methods, values of the SE and SEE for each item were highly similar, suggesting a good approximation of the estimator's standard error using Eq. (15).

Table 3 and Figure 2 present the recovery results for the item threshold parameters (b_js) in a similar manner. The proposed method that incorporates RT censoring (method 1) consistently resulted in accurate item threshold parameter estimates across conditions. On the other hand, without incorporating the RT censoring, methods 2 and 3 demonstrated consistent and significant

TABLE 3.
continued

Condition	ρ	J	N	Bias			RMSE			SE		
				M1	M2	M3	M1	M2	M3	M1	M2	M3
C3	−0.4	10	10000	−0.0004	0.0068	0.0182	0.0295	0.0303	0.0347	0.0295	0.0295	0.0295
C3	−0.4	40	1000	0.0025	0.0026	0.0100	0.0941	0.0941	0.0944	0.0941	0.0941	0.0939
C3	−0.4	40	4000	0.0019	0.0022	0.0091	0.0467	0.0467	0.0473	0.0466	0.0466	0.0464
C3	−0.4	40	10000	−0.0006	−0.0005	0.0064	0.0286	0.0286	0.0291	0.0286	0.0286	0.0284
C3	0	10	1000	0.0066	0.0066	0.0066	0.0966	0.0965	0.0964	0.0963	0.0963	0.0962
C3	0	10	4000	0.0029	0.0029	0.0028	0.0469	0.0469	0.0469	0.0468	0.0469	0.0469
C3	0	10	10000	0.0002	0.0002	0.0002	0.0300	0.0300	0.0299	0.0299	0.0299	0.0299
C3	0	40	1000	0.0020	0.0020	0.0020	0.0937	0.0937	0.0937	0.0937	0.0937	0.0937
C3	0	40	4000	0.0023	0.0023	0.0023	0.0471	0.0471	0.0471	0.0470	0.0470	0.0470
C3	0	40	10000	0.0001	0.0001	0.0001	0.0286	0.0286	0.0286	0.0286	0.0286	0.0286
C3	0.8	10	1000	0.0049	−0.0128	−0.0320	0.0965	0.0968	0.1018	0.0964	0.0960	0.0967
C3	0.8	10	4000	0.0029	−0.0147	−0.0342	0.0465	0.0486	0.0576	0.0465	0.0463	0.0463
C3	0.8	10	10000	0.0000	−0.0177	−0.0370	0.0300	0.0347	0.0475	0.0300	0.0299	0.0299
C3	0.8	40	1000	0.0031	0.0023	−0.0093	0.0967	0.0966	0.0946	0.0966	0.0966	0.0941
C3	0.8	40	4000	0.0016	0.0009	−0.0109	0.0480	0.0479	0.0482	0.0480	0.0479	0.0469
C3	0.8	40	10000	0.0000	−0.0008	−0.0132	0.0295	0.0295	0.0318	0.0295	0.0295	0.0290

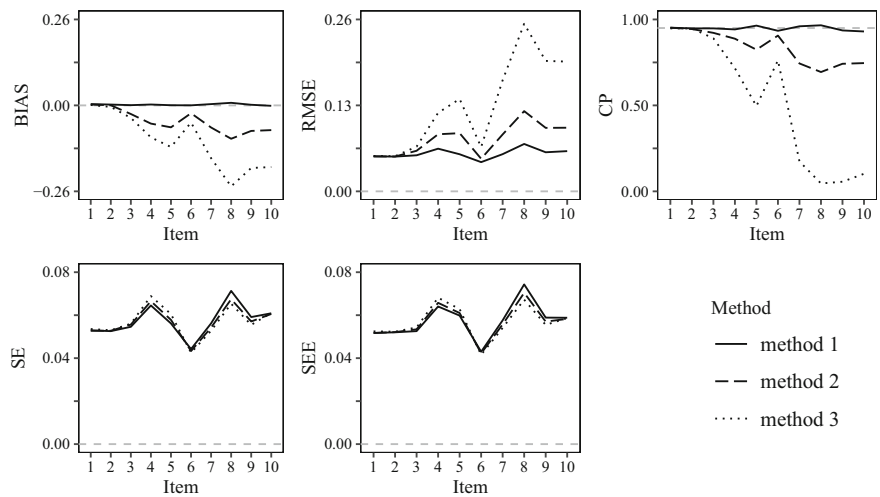

FIGURE 2.
Recovery of item threshold parameters (i.e., b_j) under C1, $J = 10$, $N = 4000$, $\rho = .8$ in terms of bias, RMSE, and 95% CP, and observed (SE) and estimated (SEE) standard errors, obtained from different estimation approaches.

bias in b estimates. The bias was larger in magnitude for method 3, which models the observed responses alone, than for method 2, which jointly models the observed responses and RTs. The direction of the bias depended on the sign of the correlation between θ and τ, and the magnitude of the bias was larger for conditions with (1) higher proportion and earlier onset of missingness, (2) fewer items, and (3) stronger association between speed and ability. Similar to a_js, the magnitude of the bias of methods 2 and 3 was larger for later items in the test, i.e., items with lower completion

TABLE 4.
Recovery of α parameters (for the last four items) under different conditions using methods 1–2 (abbreviated M1–M2).

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C1	−0.4	10	1000	0.0013	0.0627	0.0213	0.0680	0.0212	0.0224
C1	−0.4	10	4000	0.0003	0.0617	0.0108	0.0632	0.0107	0.0113
C1	−0.4	10	10000	0.0002	0.0616	0.0067	0.0622	0.0067	0.0072
C1	−0.4	40	1000	0.0020	0.0391	0.0212	0.0467	0.0211	0.0218
C1	−0.4	40	4000	0.0003	0.0375	0.0103	0.0397	0.0103	0.0105
C1	−0.4	40	10000	0.0004	0.0374	0.0066	0.0384	0.0066	0.0068
C1	0	10	1000	0.0011	0.0631	0.0215	0.0684	0.0215	0.0226
C1	0	10	4000	0.0003	0.0620	0.0108	0.0634	0.0108	0.0113
C1	0	10	10000	0.0001	0.0618	0.0067	0.0624	0.0067	0.0072
C1	0	40	1000	0.0019	0.0393	0.0211	0.0470	0.0210	0.0219
C1	0	40	4000	0.0001	0.0373	0.0103	0.0395	0.0103	0.0106
C1	0	40	10000	0.0002	0.0374	0.0067	0.0383	0.0067	0.0069
C1	0.8	10	1000	0.0014	0.0622	0.0213	0.0675	0.0213	0.0224
C1	0.8	10	4000	0.0002	0.0608	0.0107	0.0623	0.0107	0.0111
C1	0.8	10	10000	0.0001	0.0607	0.0067	0.0613	0.0067	0.0071
C1	0.8	40	1000	0.0013	0.0385	0.0213	0.0463	0.0212	0.0219
C1	0.8	40	4000	−0.0002	0.0371	0.0102	0.0393	0.0102	0.0105
C1	0.8	40	10000	0.0003	0.0373	0.0065	0.0382	0.0064	0.0066
C2	−0.4	10	1000	0.0020	0.1107	0.0297	0.1173	0.0296	0.0330
C2	−0.4	10	4000	0.0004	0.1087	0.0149	0.1104	0.0149	0.0162
C2	−0.4	10	10000	0.0003	0.1087	0.0095	0.1094	0.0094	0.0104
C2	−0.4	40	1000	0.0024	0.0508	0.0316	0.0633	0.0315	0.0330
C2	−0.4	40	4000	−0.0001	0.0480	0.0158	0.0521	0.0157	0.0161
C2	−0.4	40	10000	0.0003	0.0483	0.0102	0.0500	0.0102	0.0104
C2	0	10	1000	0.0021	0.1103	0.0298	0.1168	0.0297	0.0329
C2	0	10	4000	0.0005	0.1087	0.0150	0.1104	0.0150	0.0163
C2	0	10	10000	0.0003	0.1089	0.0095	0.1096	0.0095	0.0104
C2	0	40	1000	0.0024	0.0508	0.0319	0.0633	0.0319	0.0329
C2	0	40	4000	−0.0004	0.0481	0.0157	0.0522	0.0156	0.0160
C2	0	40	10000	0.0003	0.0482	0.0103	0.0500	0.0103	0.0106
C2	0.8	10	1000	0.0017	0.1091	0.0293	0.1155	0.0292	0.0326
C2	0.8	10	4000	0.0002	0.1074	0.0148	0.1091	0.0148	0.0159
C2	0.8	10	10000	0.0002	0.1078	0.0094	0.1085	0.0094	0.0103
C2	0.8	40	1000	0.0018	0.0497	0.0315	0.0621	0.0314	0.0325
C2	0.8	40	4000	−0.0006	0.0475	0.0157	0.0516	0.0157	0.0158
C2	0.8	40	10000	0.0002	0.0482	0.0102	0.0500	0.0102	0.0105

rate. For the last few items under condition C1 with $J = 10$, $\rho = .8$, $N = 4000$ (Fig. 2), the 95% CP of b_j estimates using methods 2 and 3 remarkably deviated from .95. In other words, when the RT censoring term is ignored, a lot of the times, the 95% confidence interval of item threshold estimates did not cover the true parameter.

2.2.2. RT model parameters For the two methods that jointly modeled responses and RTs (methods 1 and 2), results on item time discrimination (α_j) and time intensity (γ_j) parameter recovery are presented in Tables 4 and 5 and Figs. 3 and 4. Method 1 produced accurate α_j and γ_j

TABLE 4.
continued

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C3	−0.4	10	1000	0.0013	0.0194	0.0165	0.0285	0.0165	0.0164
C3	−0.4	10	4000	0.0004	0.0186	0.0081	0.0219	0.0080	0.0081
C3	−0.4	10	10000	0.0001	0.0183	0.0052	0.0200	0.0052	0.0052
C3	−0.4	40	1000	0.0008	0.0148	0.0154	0.0239	0.0154	0.0156
C3	−0.4	40	4000	0.0000	0.0140	0.0078	0.0176	0.0078	0.0078
C3	−0.4	40	10000	0.0001	0.0141	0.0050	0.0159	0.0050	0.0050
C3	0	10	1000	0.0011	0.0193	0.0165	0.0285	0.0164	0.0165
C3	0	10	4000	0.0004	0.0187	0.0080	0.0219	0.0080	0.0081
C3	0	10	10000	0.0000	0.0184	0.0052	0.0200	0.0052	0.0052
C3	0	40	1000	0.0008	0.0148	0.0154	0.0240	0.0154	0.0156
C3	0	40	4000	0.0000	0.0140	0.0078	0.0176	0.0078	0.0078
C3	0	40	10000	0.0002	0.0141	0.0050	0.0159	0.0050	0.0050
C3	0.8	10	1000	0.0010	0.0190	0.0162	0.0280	0.0162	0.0162
C3	0.8	10	4000	0.0004	0.0184	0.0080	0.0216	0.0080	0.0080
C3	0.8	10	10000	0.0000	0.0181	0.0051	0.0198	0.0051	0.0052
C3	0.8	40	1000	0.0009	0.0148	0.0155	0.0240	0.0155	0.0156
C3	0.8	40	4000	0.0000	0.0140	0.0078	0.0176	0.0078	0.0078
C3	0.8	40	10000	0.0002	0.0141	0.0050	0.0159	0.0050	0.0050

estimates across all conditions, with near-zero bias and RMSE approaching 0 as N increased. On the other hand, method 2 demonstrated a consistent trend to overestimate α_js and underestimate γ_js across all conditions. The biases were most salient when the test was short ($J = 10$), and the proportion of missingness was high (C1 or C2). The correlation between θ and τ did not affect the magnitude of the bias. From the item-specific plots under condition C1 with $J = 10$, $N = 4000$, $\rho = .8$ (Figs. 3 and 4), it can be observed that the α_j estimates from method 2 were affected for all items, regardless of serial position, and the γ_j estimates from method 2 were remarkably affected for items with any degree of missingness.

2.2.3. Structural model parameters Tables 6 and 7 report the parameter recovery results of the structural parameters, ρ and σ_τ, using methods 1 and 2. Under the proposed approach incorporating RT censoring (method 1), the ρs and σ_τs were well recovered across all conditions, with bias and RMSE approaching 0 as sample size (N) increased. On the other hand, although method 2 performed comparably in ρ and σ_τ recovery under most conditions, the bias and RMSEs were slightly larger when the proportion of missingness was higher. In particular, method 2 displayed a tendency to underestimate σ_τ, especially for shorter tests ($J = 10$) and for conditions with lower overall completion rate (C1 followed by C2).

2.2.4. Individual latent traits The RMSEs of the latent trait estimates under each simulation condition, as well as the Kendall's rank correlations between latent trait estimates produced with different methods, are presented in Table 8. In terms of latent ability (θ) estimation, it could be observed that the three methods barely differed when the test was long ($J = 40$) or when the latent traits were uncorrelated ($\rho = 0$). Under these conditions, the RMSEs for θ with the three methods were highly similar, and this was especially the case for methods 1 and 2, with Kendall's correlation in the estimated θs close to 1. When the test length was short and the correlation

TABLE 5.
Recovery of γ parameters (for the last four items) under different conditions using methods 1–2 (abbreviated M1–M2).

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C1	−0.4	10	1000	−0.0022	−0.3232	0.0689	0.3319	0.0689	0.0662
C1	−0.4	10	4000	−0.0007	−0.3231	0.0339	0.3252	0.0339	0.0322
C1	−0.4	10	10000	−0.0007	−0.3228	0.0220	0.3237	0.0220	0.0211
C1	−0.4	40	1000	−0.0013	−0.1608	0.0707	0.1827	0.0707	0.0684
C1	−0.4	40	4000	−0.0010	−0.1617	0.0348	0.1674	0.0348	0.0337
C1	−0.4	40	10000	−0.0004	−0.1604	0.0220	0.1629	0.0220	0.0212
C1	0	10	1000	−0.0016	−0.3260	0.0689	0.3345	0.0690	0.0663
C1	0	10	4000	−0.0003	−0.3256	0.0341	0.3278	0.0341	0.0327
C1	0	10	10000	−0.0005	−0.3255	0.0218	0.3264	0.0218	0.0210
C1	0	40	1000	−0.0010	−0.1615	0.0701	0.1827	0.0701	0.0673
C1	0	40	4000	−0.0011	−0.1621	0.0347	0.1677	0.0347	0.0334
C1	0	40	10000	−0.0002	−0.1605	0.0220	0.1630	0.0220	0.0213
C1	0.8	10	1000	−0.0020	−0.3124	0.0689	0.3215	0.0689	0.0662
C1	0.8	10	4000	0.0004	−0.3110	0.0335	0.3133	0.0335	0.0325
C1	0.8	10	10000	−0.0004	−0.3115	0.0219	0.3125	0.0219	0.0210
C1	0.8	40	1000	−0.0020	−0.1617	0.0705	0.1830	0.0705	0.0683
C1	0.8	40	4000	−0.0011	−0.1619	0.0349	0.1674	0.0349	0.0332
C1	0.8	40	10000	0.0003	−0.1598	0.0220	0.1623	0.0220	0.0214
C2	−0.4	10	1000	0.0002	−0.2248	0.0403	0.2302	0.0403	0.0393
C2	−0.4	10	4000	−0.0003	−0.2252	0.0199	0.2265	0.0198	0.0189
C2	−0.4	10	10000	−0.0003	−0.2253	0.0125	0.2259	0.0125	0.0121
C2	−0.4	40	1000	−0.0007	−0.1168	0.0456	0.1282	0.0456	0.0441
C2	−0.4	40	4000	−0.0008	−0.1171	0.0221	0.1201	0.0221	0.0218
C2	−0.4	40	10000	−0.0003	−0.1163	0.0141	0.1175	0.0141	0.0138
C2	0	10	1000	0.0002	−0.2250	0.0403	0.2302	0.0403	0.0388
C2	0	10	4000	−0.0004	−0.2253	0.0201	0.2266	0.0201	0.0190
C2	0	10	10000	−0.0004	−0.2257	0.0125	0.2262	0.0125	0.0124
C2	0	40	1000	−0.0008	−0.1172	0.0446	0.1284	0.0446	0.0435
C2	0	40	4000	−0.0001	−0.1168	0.0225	0.1199	0.0225	0.0221
C2	0	40	10000	−0.0002	−0.1162	0.0139	0.1174	0.0139	0.0135
C2	0.8	10	1000	−0.0003	−0.2239	0.0390	0.2291	0.0391	0.0382
C2	0.8	10	4000	−0.0002	−0.2238	0.0197	0.2251	0.0196	0.0191
C2	0.8	10	10000	−0.0002	−0.2242	0.0126	0.2248	0.0126	0.0123
C2	0.8	40	1000	−0.0024	−0.1179	0.0460	0.1293	0.0459	0.0446
C2	0.8	40	4000	0.0000	−0.1159	0.0225	0.1192	0.0225	0.0221

between θ and τ increased, however, the three approaches start to visibly differ in their ranking of individual abilities, with the Kendall's correlation between methods 1 and 2 dropping to .94 (and .81 between methods 1 and 3) under condition C1 with $\rho = .8$. Under these scenarios, the RMSE of θ was lowest with method 1, followed closely by method 2, suggesting a slight advantage of the proposed method in latent ability recovery.

The difference between methods 1 and 2 was more apparent in the estimates of the latent speeds (τs), especially when the test was short. The Kendall's rank correlation between the τ estimates produced with the two methods ranged between .84 and .96 when $J = 10$. Across all

TABLE 5.
continued

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C2	0.8	40	10000	0.0002	−0.1157	0.0142	0.1171	0.0142	0.0139
C3	−0.4	10	1000	−0.0005	−0.0741	0.0553	0.0995	0.0553	0.0548
C3	−0.4	10	4000	−0.0003	−0.0743	0.0270	0.0818	0.0270	0.0266
C3	−0.4	10	10000	0.0001	−0.0740	0.0174	0.0775	0.0174	0.0172
C3	−0.4	40	1000	-0.0010	−0.0507	0.0540	0.0829	0.0540	0.0539
C3	−0.4	40	4000	-0.0011	−0.0510	0.0275	0.0629	0.0275	0.0272
C3	−0.4	40	10000	-0.0004	−0.0502	0.0173	0.0563	0.0173	0.0171
C3	0	10	1000	−0.0002	−0.0745	0.0549	0.0994	0.0550	0.0545
C3	0	10	4000	0.0000	−0.0747	0.0267	0.0820	0.0267	0.0265
C3	0	10	10000	0.0000	−0.0747	0.0174	0.0781	0.0174	0.0172
C3	0	40	1000	−0.0008	−0.0505	0.0539	0.0827	0.0539	0.0536
C3	0	40	4000	−0.0006	−0.0506	0.0276	0.0628	0.0276	0.0274
C3	0	40	10000	−0.0003	−0.0501	0.0172	0.0562	0.0172	0.0170
C3	0.8	10	1000	−0.0005	−0.0719	0.0545	0.0974	0.0545	0.0542
C3	0.8	10	4000	0.0004	−0.0715	0.0267	0.0794	0.0266	0.0264
C3	0.8	10	10000	0.0000	−0.0719	0.0174	0.0755	0.0174	0.0171
C3	0.8	40	1000	−0.0026	−0.0522	0.0552	0.0842	0.0551	0.0545
C3	0.8	40	4000	−0.0005	−0.0504	0.0280	0.0629	0.0280	0.0277
C3	0.8	40	10000	−0.0003	−0.0501	0.0176	0.0564	0.0176	0.0173

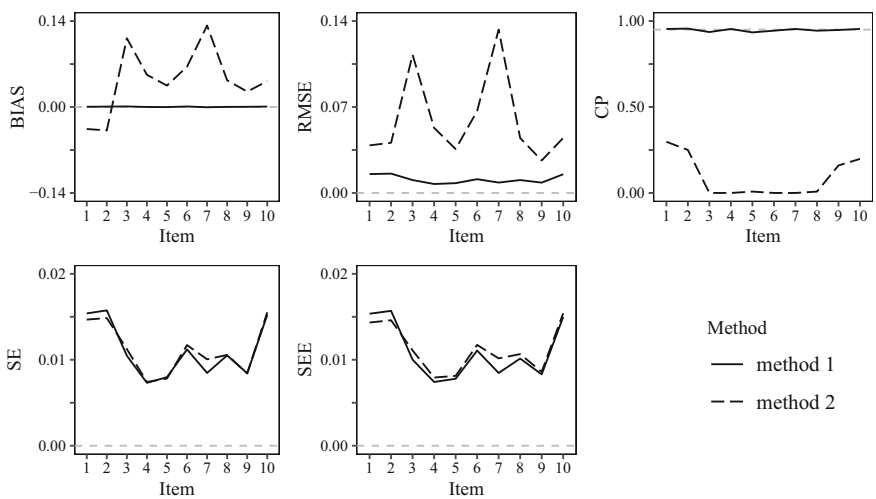

FIGURE 3.
Recovery of item time discrimination parameters (i.e., α_j) under C1, $J = 10$, $N = 4000$, $\rho = .8$ in terms of bias, RMSE, and 95% CP, as well as the observed (SE) and estimated (SEE) standard errors, obtained from different estimation approaches.

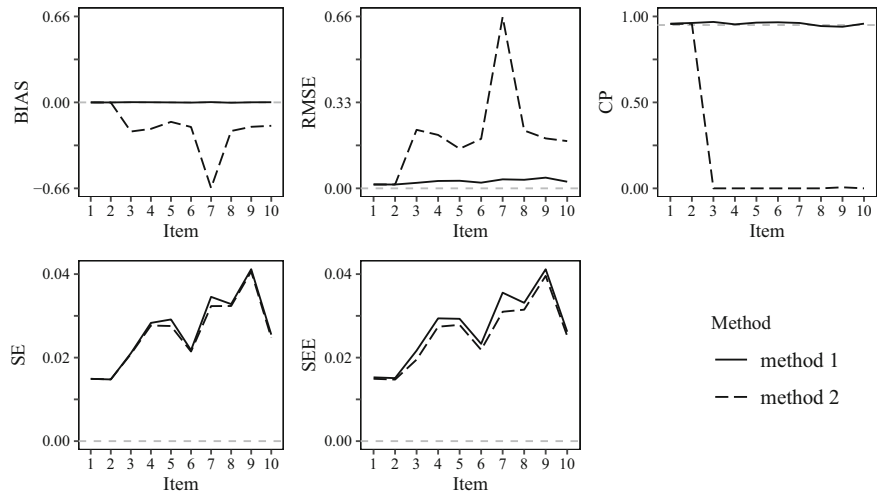

FIGURE 4.
Recovery of item time intensity parameters (i.e., γ_j) under C1 with $J = 10$, $N = 4000$, $\rho = .8$ in terms of bias, RMSE, and 95% CP, and observed (SE) and estimated (SEE) standard errors, obtained from different estimation approaches.

conditions, method 1 produced lower, if not equal, RMSE in τ estimates compared to method 2, suggesting better latent speed recovery in the presence of missingness due to RT censoring.

3. Application: PISA 2018 Science Exam

The three methods were applied to the data from the computer-based version of the PISA 2018 Science Test. In PISA 2018, test takers were randomly routed to different clusters of science items, and each cluster of items was administered as two parts. Each part was expected to be completed within 30 minutes. Instead of imposing a 30-minute time limit on each part, a total time limit of 60 minutes was imposed for the two parts combined (OECD 2021).

For the current study, the response and RT data on items administered in the first position of science trend form number 23 (i.e., S05 cluster in form 23) were analyzed. Following Pohl et al. (2019), an artificial time limit of 30 minutes was imposed for items in the first part, and any responses and RTs to the first part items that exceeded the 30-minute time limit were treated as not reached. To ensure a certain degree of homogeneity in test-taker characteristics, the sample was chosen to be the test takers from 16 countries whose average science scores are between 490 and 508 (Schleicher 2019). Note that NRI was not the only type of missingness that was present in the data. For the purpose of the current study, examinees who manifested other types of missingness, such as omissions and early quitting, were removed from further analyses. Some examinees may also demonstrate rapid-guessing behavior by responding quickly to items without engaging in cognitive problem-solving processes. Treating such rapid-guessing behavior as solution behavior in modeling and scoring may jeopardize the validity of the test scores (Wise and Kingsbury 2016). Following Wise and Ma (2012), an item-wise threshold for rapid guessing was set at 10% of its average observed RTs, with a maximum threshold value of 10 seconds. Examinees' responses were classified as rapid guesses when the RTs were less than the item threshold. The examinees with rapid guessing behavior were further removed from the data set. This resulted in a total of $N = 2335$ examinees.

$J = 20$ items were administered in science cluster S05. For simplicity, polytomous responses were recoded to dichotomous scores based on whether the examinee had received the highest

TABLE 6.
Recovery of the correlation of θ and τ with method 1 (M1), method 2 (M2), under different conditions.

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C1	−0.4	10	1000	0.0000	−0.0187	0.0443	0.0518	0.0443	0.0484
C1	−0.4	10	4000	0.0018	−0.0167	0.0246	0.0315	0.0246	0.0268
C1	−0.4	10	10000	0.0007	−0.0177	0.0156	0.0246	0.0156	0.0171
C1	−0.4	40	1000	0.0005	0.0004	0.0295	0.0294	0.0295	0.0295
C1	−0.4	40	4000	0.0005	0.0005	0.0150	0.0150	0.0150	0.0150
C1	−0.4	40	10000	0.0004	0.0003	0.0095	0.0095	0.0095	0.0096
C1	0	10	1000	0.0024	0.0020	0.0511	0.0551	0.0511	0.0551
C1	0	10	4000	0.0008	0.0017	0.0260	0.0279	0.0261	0.0279
C1	0	10	10000	0.0004	0.0005	0.0169	0.0181	0.0170	0.0181
C1	0	40	1000	0.0008	0.0009	0.0344	0.0344	0.0344	0.0345
C1	0	40	4000	0.0005	0.0005	0.0178	0.0178	0.0178	0.0178
C1	0	40	10000	−0.0002	−0.0002	0.0111	0.0111	0.0112	0.0111
C1	0.8	10	1000	0.0017	0.0378	0.0357	0.0546	0.0357	0.0394
C1	0.8	10	4000	−0.0011	0.0353	0.0163	0.0397	0.0163	0.0182
C1	0.8	10	10000	−0.0005	0.0354	0.0108	0.0373	0.0108	0.0120
C1	0.8	40	1000	−0.0001	0.0001	0.0145	0.0145	0.0145	0.0145
C1	0.8	40	4000	−0.0002	0.0001	0.0074	0.0074	0.0074	0.0074
C1	0.8	40	10000	−0.0007	−0.0004	0.0048	0.0048	0.0048	0.0048
C2	−0.4	10	1000	−0.0002	−0.0016	0.0345	0.0349	0.0345	0.0349
C2	−0.4	10	4000	0.0011	−0.0004	0.0187	0.0188	0.0186	0.0188
C2	−0.4	10	10000	0.0007	−0.0008	0.0118	0.0119	0.0118	0.0118
C2	−0.4	40	1000	0.0002	0.0002	0.0293	0.0293	0.0294	0.0294
C2	−0.4	40	4000	0.0007	0.0007	0.0150	0.0151	0.0150	0.0151
C2	−0.4	40	10000	0.0005	0.0005	0.0096	0.0096	0.0096	0.0096
C2	0	10	1000	0.0005	0.0007	0.0406	0.0412	0.0407	0.0413
C2	0	10	4000	0.0012	0.0012	0.0204	0.0207	0.0204	0.0207
C2	0	10	10000	0.0001	0.0000	0.0134	0.0135	0.0134	0.0135
C2	0	40	1000	0.0008	0.0009	0.0345	0.0345	0.0346	0.0346
C2	0	40	4000	0.0005	0.0006	0.0179	0.0179	0.0179	0.0179
C2	0	40	10000	−0.0002	−0.0002	0.0113	0.0113	0.0113	0.0113
C2	0.8	10	1000	−0.0002	0.0035	0.0221	0.0225	0.0221	0.0222
C2	0.8	10	4000	−0.0002	0.0036	0.0101	0.0109	0.0101	0.0103
C2	0.8	10	10000	−0.0004	0.0033	0.0070	0.0077	0.0070	0.0070
C2	0.8	40	1000	−0.0001	0.0001	0.0144	0.0144	0.0144	0.0144
C2	0.8	40	4000	−0.0004	−0.0001	0.0074	0.0074	0.0074	0.0074
C2	0.8	40	10000	−0.0009	−0.0007	0.0048	0.0048	0.0047	0.0047
C3	−0.4	10	1000	−0.0010	−0.0024	0.0402	0.0407	0.0403	0.0406
C3	−0.4	10	4000	0.0018	0.0004	0.0223	0.0224	0.0223	0.0225
C3	−0.4	10	10000	0.0008	−0.0008	0.0137	0.0138	0.0136	0.0138
C3	−0.4	40	1000	0.0006	0.0006	0.0287	0.0287	0.0287	0.0287
C3	−0.4	40	4000	0.0005	0.0005	0.0148	0.0148	0.0148	0.0148
C3	−0.4	40	10000	0.0003	0.0003	0.0092	0.0092	0.0092	0.0092
C3	0	10	1000	0.0000	0.0000	0.0460	0.0464	0.0460	0.0465
C3	0	10	4000	0.0020	0.0020	0.0238	0.0240	0.0238	0.0240
C3	0	10	10000	0.0003	0.0001	0.0153	0.0155	0.0153	0.0155

TABLE 6.
continued

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C3	0	40	1000	0.0011	0.0010	0.0339	0.0339	0.0339	0.0339
C3	0	40	4000	0.0007	0.0007	0.0173	0.0173	0.0173	0.0173
C3	0	40	10000	−0.0002	−0.0002	0.0109	0.0109	0.0109	0.0109
C3	0.8	10	1000	−0.0004	0.0036	0.0318	0.0324	0.0318	0.0322
C3	0.8	10	4000	−0.0005	0.0035	0.0146	0.0152	0.0146	0.0148
C3	0.8	10	10000	−0.0003	0.0036	0.0093	0.0101	0.0093	0.0095
C3	0.8	40	1000	0.0001	0.0001	0.0141	0.0141	0.0141	0.0141
C3	0.8	40	4000	−0.0002	−0.0001	0.0072	0.0072	0.0072	0.0072
C3	0.8	40	10000	−0.0006	−0.0006	0.0046	0.0046	0.0046	0.0046

possible score on the item. Table 9 presents summary statistics of the items calculated based on the 2335 examinees' observed responses and RTs. Approximately 43.7% of examinees completed all 20 items within 30 minutes, and the earliest onset of NRI due to the 30-minute time limit was at the 4th item.

3.1. Results

Both method 1 and method 2 identified a slight negative overall correlation between latent speed and ability, with $\hat{\rho} = -.169$ using the joint model with censoring (method 1), and $\hat{\rho} = -.175$ using the joint model without censoring (method 2). The estimated standard deviations of τ, $\hat{\sigma}_\tau$, using methods 1 and 2, were .245 and .236, respectively. Consistent with the results from the simulation studies, when the two latent traits are negatively correlated, the joint model without censoring appeared to produce slightly lower $\hat{\rho}$ and $\hat{\sigma}_\tau$ compared to the joint model with censoring.

The estimated item parameters using different methods, together with the standard error estimates (in the parentheses), are shown in Table 10. In addition, the differences in the estimated item parameters using the three methods are given in Table 11. Because the estimated correlation between θ and τ was relatively weak, one would expect the differences in the item parameter and latent ability estimates produced with the three approaches to be small. On the estimation of the a parameters, the difference between any two methods was relatively small. For b parameters, especially for items at the end of the test, one can observe a slight but consistent tendency for method 3 (without incorporating RTs) to produce larger point estimates compared to method 1 and method 2. However, the difference across these methods was below 1 $\hat{\text{SE}}$ for all items. As for the RT model item parameters, comparing methods 1 and 2, the α_j estimate produced by method 2 was more than 1 $\hat{\text{SE}}$ lower than that from method 1 for item 15, and the γ_j estimates from method 2 were more than 1 $\hat{\text{SE}}$ lower than those from method 1 for items 14–20. This was consistent with the simulation results, where the joint model without censoring resulted in lower time intensity estimates compared to the joint model with censoring, particularly for items at the end of the test.

The examinees' latent trait estimates produced from different methods are plotted in Fig. 5. On each subplot, the gray dots represent individuals who have completed all 20 items within the first 30 minutes, and the black dots represent individuals with at least 1 not-reached item up to 30 minutes. Comparing the latent ability estimates produced by methods 1 and 2, it is observed that

TABLE 7.
Recovery of standard deviation of τ with method 1 (M1), method 2 (M2), under different conditions.

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C1	−0.4	10	1000	−0.0008	-0.0976	0.0205	0.1000	0.0205	0.0216
C1	−0.4	10	4000	−0.0008	-0.0978	0.0103	0.0984	0.0103	0.0108
C1	−0.4	10	10000	−0.0006	-0.0973	0.0060	0.0975	0.0060	0.0063
C1	−0.4	40	1000	−0.0008	-0.0091	0.0120	0.0150	0.0120	0.0119
C1	−0.4	40	4000	−0.0005	-0.0089	0.0059	0.0107	0.0059	0.0058
C1	−0.4	40	10000	−0.0001	-0.0085	0.0040	0.0094	0.0040	0.0040
C1	0	10	1000	−0.0009	−0.0988	0.0207	0.1011	0.0207	0.0216
C1	0	10	4000	−0.0006	−0.0985	0.0101	0.0991	0.0101	0.0107
C1	0	10	10000	−0.0006	−0.0983	0.0062	0.0986	0.0062	0.0067
C1	0	40	1000	−0.0005	−0.0089	0.0118	0.0147	0.0118	0.0117
C1	0	40	4000	−0.0003	−0.0087	0.0060	0.0106	0.0060	0.0059
C1	0	40	10000	0.0000	−0.0085	0.0040	0.0093	0.0040	0.0040
C1	0.8	10	1000	−0.0010	−0.0938	0.0209	0.0962	0.0209	0.0213
C1	0.8	10	4000	0.0001	−0.0932	0.0101	0.0938	0.0101	0.0106
C1	0.8	10	10000	−0.0005	−0.0934	0.0067	0.0937	0.0067	0.0070
C1	0.8	40	1000	−0.0001	−0.0084	0.0121	0.0146	0.0122	0.0120
C1	0.8	40	4000	−0.0001	−0.0084	0.0062	0.0104	0.0062	0.0062
C1	0.8	40	10000	−0.0002	−0.0085	0.0040	0.0094	0.0040	0.0039
C2	−0.4	10	1000	−0.0004	−0.0269	0.0138	0.0302	0.0138	0.0138
C2	−0.4	10	4000	−0.0008	−0.0274	0.0068	0.0282	0.0067	0.0067
C2	−0.4	10	10000	−0.0003	−0.0269	0.0042	0.0272	0.0042	0.0042
C2	−0.4	40	1000	−0.0006	−0.0106	0.0119	0.0159	0.0119	0.0118
C2	−0.4	40	4000	−0.0005	−0.0105	0.0059	0.0120	0.0059	0.0058
C2	−0.4	40	10000	0.0000	−0.0101	0.0040	0.0108	0.0040	0.0040
C2	0	10	1000	−0.0006	−0.0271	0.0138	0.0304	0.0138	0.0138
C2	0	10	4000	−0.0006	−0.0272	0.0067	0.0280	0.0067	0.0068
C2	0	10	10000	−0.0003	−0.0269	0.0042	0.0273	0.0042	0.0042
C2	0	40	1000	−0.0004	−0.0105	0.0119	0.0158	0.0119	0.0118
C2	0	40	4000	−0.0002	−0.0103	0.0060	0.0119	0.0060	0.0059
C2	0	40	10000	0.0001	−0.0100	0.0040	0.0108	0.0040	0.0040
C2	0.8	10	1000	−0.0004	−0.0266	0.0135	0.0299	0.0136	0.0136
C2	0.8	10	4000	−0.0002	−0.0265	0.0069	0.0274	0.0069	0.0070
C2	0.8	10	10000	−0.0004	−0.0268	0.0045	0.0272	0.0045	0.0045
C2	0.8	40	1000	0.0001	−0.0099	0.0123	0.0157	0.0123	0.0122
C2	0.8	40	4000	0.0001	−0.0098	0.0062	0.0116	0.0063	0.0062
C2	0.8	40	10000	−0.0001	−0.0100	0.0039	0.0107	0.0039	0.0039
C3	−0.4	10	1000	−0.0002	−0.0236	0.0189	0.0302	0.0189	0.0188
C3	−0.4	10	4000	−0.0008	−0.0242	0.0093	0.0259	0.0092	0.0092
C3	−0.4	10	10000	−0.0004	−0.0239	0.0055	0.0245	0.0054	0.0054
C3	−0.4	40	1000	−0.0007	−0.0026	0.0119	0.0121	0.0119	0.0118
C3	−0.4	40	4000	−0.0005	−0.0023	0.0058	0.0062	0.0058	0.0058
C3	−0.4	40	10000	−0.0001	−0.0020	0.0040	0.0044	0.0040	0.0040
C3	0	10	1000	−0.0004	−0.0241	0.0187	0.0305	0.0188	0.0187
C3	0	10	4000	−0.0003	−0.0239	0.0093	0.0256	0.0093	0.0091

TABLE 7.
continued

Condition	ρ	J	N	Bias		RMSE		SE	
				M1	M2	M1	M2	M1	M2
C3	0	10	10000	−0.0005	−0.0241	0.0056	0.0248	0.0056	0.0056
C3	0	40	1000	−0.0006	−0.0025	0.0118	0.0120	0.0118	0.0118
C3	0	40	4000	−0.0003	−0.0022	0.0060	0.0063	0.0060	0.0060
C3	0	40	10000	−0.0001	−0.0020	0.0040	0.0044	0.0040	0.0040
C3	0.8	10	1000	−0.0006	−0.0233	0.0191	0.0299	0.0191	0.0188
C3	0.8	10	4000	0.0004	−0.0223	0.0093	0.0242	0.0093	0.0092
C3	0.8	10	10000	−0.0004	−0.0231	0.0061	0.0239	0.0061	0.0060
C3	0.8	40	1000	0.0001	−0.0018	0.0121	0.0122	0.0121	0.0121
C3	0.8	40	4000	−0.0001	−0.0019	0.0061	0.0064	0.0061	0.0061
C3	0.8	40	10000	−0.0002	−0.0021	0.0039	0.0044	0.0039	0.0039

TABLE 8.
Recovery of the examinee latent traits under different conditions using methods 1–3 (abbreviated M1–M3). COR represents the Kendall's rank correlation coefficient of the estimates of latent parameters with different methods

Condition	ρ	J	N	RMSE					COR			
				θ			τ		θ			τ
				M1	M2	M3	M1	M2	M1&M2	M1&M3	M2&M3	M1&M2
C1	−0.4	10	1000	0.5747	0.5763	0.5888	0.3155	0.3412	0.9719	0.9095	0.9223	0.8408
C1	−0.4	10	4000	0.5734	0.5748	0.5870	0.3153	0.3411	0.9721	0.9105	0.9232	0.8408
C1	−0.4	10	10000	0.5723	0.5738	0.5862	0.3152	0.3411	0.9721	0.9103	0.9231	0.8411
C1	−0.4	40	1000	0.3245	0.3245	0.3294	0.1150	0.1157	0.9993	0.9690	0.9691	0.9893
C1	−0.4	40	4000	0.3226	0.3226	0.3276	0.1137	0.1146	0.9993	0.9690	0.9691	0.9893
C1	−0.4	40	10000	0.3217	0.3217	0.3267	0.1136	0.1145	0.9993	0.9689	0.9691	0.9893
C1	0	10	1000	0.5871	0.5871	0.5869	0.3229	0.3495	0.9948	0.9878	0.9889	0.8351
C1	0	10	4000	0.5848	0.5848	0.5848	0.3225	0.3493	0.9969	0.9922	0.9927	0.8353
C1	0	10	10000	0.5842	0.5842	0.5842	0.3226	0.3494	0.9974	0.9937	0.9940	0.8356
C1	0	40	1000	0.3269	0.3269	0.3269	0.1153	0.1161	0.9999	0.9980	0.9980	0.9892
C1	0	40	4000	0.3248	0.3248	0.3248	0.1142	0.1151	1.0000	0.9990	0.9990	0.9892
C1	0	40	10000	0.3242	0.3242	0.3242	0.1141	0.1149	1.0000	0.9994	0.9994	0.9892
C1	0.8	10	1000	0.5348	0.5430	0.6024	0.2849	0.3078	0.9386	0.8098	0.8324	0.8667
C1	0.8	10	4000	0.5319	0.5402	0.6010	0.2838	0.3067	0.9389	0.8106	0.8331	0.8662
C1	0.8	10	10000	0.5321	0.5403	0.6006	0.2839	0.3069	0.9389	0.8112	0.8335	0.8667
C1	0.8	40	1000	0.3009	0.3009	0.3319	0.1118	0.1125	0.9978	0.9222	0.9225	0.9899
C1	0.8	40	4000	0.2986	0.2987	0.3298	0.1107	0.1114	0.9978	0.9223	0.9226	0.9899
C1	0.8	40	10000	0.2981	0.2981	0.3295	0.1105	0.1112	0.9978	0.9224	0.9227	0.9899
C2	−0.4	10	1000	0.5152	0.5154	0.5285	0.2005	0.2057	0.9935	0.9235	0.9248	0.9522
C2	−0.4	10	4000	0.5137	0.5138	0.5270	0.2001	0.2055	0.9935	0.9241	0.9254	0.9522
C2	−0.4	10	10000	0.5129	0.5130	0.5263	0.2000	0.2054	0.9935	0.9240	0.9253	0.9522
C2	−0.4	40	1000	0.3282	0.3282	0.3337	0.1117	0.1127	0.9991	0.9664	0.9667	0.9890
C2	−0.4	40	4000	0.3262	0.3262	0.3319	0.1102	0.1114	0.9991	0.9665	0.9667	0.9890

TABLE 8.
continued

Condition	ρ	J	N	RMSE					COR			
				θ			τ		θ			τ
				M1	M2	M3	M1	M2	M1&M2	M1&M3	M2&M3	M1&M2
C2	−0.4	40	10000	0.3254	0.3254	0.3312	0.1100	0.1112	0.9991	0.9664	0.9666	0.9891
C2	0	10	1000	0.5254	0.5254	0.5253	0.2026	0.2079	0.9989	0.9928	0.9928	0.9513
C2	0	10	4000	0.5234	0.5234	0.5233	0.2023	0.2077	0.9993	0.9956	0.9956	0.9512
C2	0	10	10000	0.5229	0.5229	0.5229	0.2021	0.2076	0.9994	0.9964	0.9965	0.9512
C2	0	40	1000	0.3301	0.3301	0.3300	0.1120	0.1130	0.9999	0.9978	0.9979	0.9889
C2	0	40	4000	0.3281	0.3281	0.3281	0.1106	0.1118	1.0000	0.9989	0.9989	0.9889
C2	0	40	10000	0.3276	0.3276	0.3276	0.1104	0.1116	1.0000	0.9993	0.9993	0.9890
C2	0.8	10	1000	0.4543	0.4553	0.5289	0.1897	0.1945	0.9825	0.8262	0.8291	0.9576
C2	0.8	10	4000	0.4522	0.4532	0.5274	0.1891	0.1939	0.9825	0.8263	0.8292	0.9576
C2	0.8	10	10000	0.4520	0.4530	0.5269	0.1890	0.1939	0.9825	0.8266	0.8296	0.9575
C2	0.8	40	1000	0.3030	0.3029	0.3369	0.1091	0.1100	0.9974	0.9181	0.9186	0.9896
C2	0.8	40	4000	0.3003	0.3003	0.3349	0.1075	0.1084	0.9974	0.9182	0.9187	0.9896
C2	0.8	40	10000	0.2996	0.2998	0.3345	0.1072	0.1082	0.9974	0.9183	0.9188	0.9896
C3	−0.4	10	1000	0.5098	0.5103	0.5188	0.3009	0.3056	0.9917	0.9386	0.9418	0.9523
C3	−0.4	10	4000	0.5084	0.5088	0.5174	0.3006	0.3054	0.9919	0.9393	0.9424	0.9526
C3	−0.4	10	10000	0.5076	0.5081	0.5167	0.3005	0.3054	0.9919	0.9392	0.9423	0.9525
C3	−0.4	40	1000	0.3043	0.3044	0.3072	0.1117	0.1118	0.9998	0.9766	0.9766	0.9965
C3	−0.4	40	4000	0.3022	0.3022	0.3052	0.1104	0.1105	0.9998	0.9766	0.9766	0.9965
C3	−0.4	40	10000	0.3014	0.3014	0.3044	0.1103	0.1104	0.9998	0.9766	0.9766	0.9965
C3	0	10	1000	0.5179	0.5179	0.5178	0.3083	0.3132	0.9985	0.9917	0.9920	0.9498
C3	0	10	4000	0.5158	0.5158	0.5158	0.3080	0.3130	0.9989	0.9942	0.9943	0.9499
C3	0	10	10000	0.5155	0.5155	0.5155	0.3079	0.3130	0.9991	0.9951	0.9952	0.9499
C3	0	40	1000	0.3068	0.3068	0.3068	0.1120	0.1121	1.0000	0.9986	0.9986	0.9964
C3	0	40	4000	0.3047	0.3047	0.3047	0.1108	0.1109	1.0000	0.9993	0.9993	0.9965
C3	0	40	10000	0.3041	0.3041	0.3041	0.1107	0.1108	1.0000	0.9995	0.9995	0.9965
C3	0.8	10	1000	0.4760	0.4786	0.5228	0.2674	0.2714	0.9822	0.8652	0.8706	0.9622
C3	0.8	10	4000	0.4739	0.4765	0.5211	0.2666	0.2705	0.9824	0.8652	0.8706	0.9625
C3	0.8	10	10000	0.4737	0.4762	0.5206	0.2666	0.2706	0.9824	0.8655	0.8709	0.9625
C3	0.8	40	1000	0.2860	0.2860	0.3076	0.1088	0.1089	0.9993	0.9361	0.9361	0.9967
C3	0.8	40	4000	0.2834	0.2834	0.3054	0.1075	0.1076	0.9993	0.9362	0.9362	0.9967
C3	0.8	40	10000	0.2827	0.2827	0.3050	0.1074	0.1075	0.9993	0.9363	0.9363	0.9967

for some individuals with NRIs (black dots), their ability rankings appeared to be slightly lower when the censoring was not explicitly modeled. This tendency was more apparent comparing methods 1 and 3. For the latent speeds produced by methods 1 and 2, it could be seen that for individuals without NRI (gray dots), the estimates produced by the two methods almost lie on a straight line. However, for individuals with NRIs (black dots), their speed estimates were relatively higher when the censoring was not explicitly modeled.

Model fit of methods 1 and 2 was further evaluated by comparing the empirical cumulative distribution functions (ECDFs) of cumulative observed RTs (i.e., $\sum_{j=1}^{J} \tilde{T}_{ij}, i = 1, \ldots, N$) with the model-implied ones. Figure 6 presents the P–P plot (empirical vs. theoretical CDF) and the Q–Q plot (empirical vs. theoretical quantile functions), where the solid line represents method 1 and the dashed line represents method 2. For the most part, the two methods performed comparably and showed very high agreement with the observed cumulative RTs except at the low end of the

TABLE 9.
Descriptive statistics of items in PISA 2018 science cluster S05 in form 23.

Item ID	N(Reached)	P(correct)	Mean Log RT		SD Log RT	
			Correct	Incorrect	Correct	Incorrect
CS498Q02S	2335 (1.000)	0.406	0.260	0.321	0.398	0.420
CS498Q03S	2335 (1.000)	0.487	−0.071	0.061	0.465	0.482
DS498Q04C	2334 (1.000)	0.689	0.907	0.586	0.448	0.580
DS514Q02C	2331 (0.998)	0.921	0.633	0.737	0.469	0.509
DS514Q03C	2327 (0.997)	0.484	0.757	0.774	0.481	0.524
DS514Q04C	2320 (0.994)	0.651	0.691	0.775	0.350	0.424
CS605Q01S	2309 (0.989)	0.604	0.529	0.494	0.419	0.493
CS605Q02S	2303 (0.986)	0.472	0.234	0.008	0.447	0.553
CS605Q03S	2281 (0.977)	0.637	0.340	0.237	0.486	0.619
DS605Q04C	2245 (0.961)	0.609	0.638	0.273	0.424	0.579
CS646Q01S	2210 (0.946)	0.912	0.060	0.206	0.354	0.553
CS646Q02S	2153 (0.922)	0.616	0.296	0.133	0.337	0.434
CS646Q03S	2089 (0.895)	0.799	0.197	−0.035	0.364	0.469
DS646Q04C	1883 (0.806)	0.363	0.943	0.521	0.338	0.508
DS646Q05C	1691 (0.724)	0.175	0.626	0.203	0.336	0.468
CS620Q01S	1606 (0.688)	0.894	−0.433	−0.416	0.357	0.487
CS620Q02S	1444 (0.618)	0.497	0.188	0.076	0.288	0.416
CS645Q01S	1304 (0.558)	0.527	−0.000	−0.132	0.411	0.480
CS645Q03S	1216 (0.521)	0.638	−0.758	−0.812	0.474	0.638
DS645Q04C	1020 (0.437)	0.661	0.189	0.080	0.357	0.491
Mean	1987 (0.851)	0.605	0.311	0.205	0.400	0.506

Notes. N(Reached) denotes the number and percentage of examinees (in the parentheses) who have reached a particular item. P(correct) denotes the proportion of responded examinees who have answered a question correctly. Mean log RT and SD log RT denote the mean and standard deviation of the natural logarithm of the observed response time on the item, grouped by examinees' response accuracy (i.e., correct or incorrect)

cumulative RTs (those with RT$\leq$ 15 minutes). However, the points produced under method 1 were slightly closer to the 45° line than those under method 2, providing evidence that method 1 fitted the data slightly better.

Moreover, Fig. 7 plots the densities of cumulative observed RTs (i.e., $\sum_{j=1}^{J} \tilde{T}_{ij}, i = 1, \ldots, N$) and model-implied ones based on kernel density estimation method. On this figure, the black solid line represents the density from observed RTs, and the red dashed line and blue long dash line, respectively, represent the density predicted by methods 1 and 2. On the left part of this figure, methods 1 and 2 perform similarly, and they have little difference with the observed one. However, on the right part of this figure, the densities obtained by method 1 and the observed one are almost coincident, and they have obvious difference with that produced by method 2, suggesting that method 1 fitted the data better.

4. Discussions

4.1. Summary of Findings

The present study proposes a way to model missingness due to not-reached items in computer-based tests using RT censoring. The time censoring mechanism was directly incorporated into the joint likelihood given the observed responses and RTs. Simulation results showed that the proposed

TABLE 10.
Estimated item parameters and standard error estimates (in the parentheses) of the PISA 2018 Science items, using different approaches.

Item ID	a			α	
	M1	M2	M3	M1	M2
CS498Q02S	0.7556 (0.0612)	0.7549 (0.0611)	0.7677 (0.0614)	2.8739 (0.0451)	2.8759 (0.0452)
CS498Q03S	0.1798 (0.0485)	0.1798 (0.0485)	0.1876 (0.0486)	2.4106 (0.0371)	2.4057 (0.0370)
DS498Q04C	1.1003 (0.0741)	1.1020 (0.0742)	1.0855 (0.0737)	2.3701 (0.0365)	2.3645 (0.0364)
DS514Q02C	1.0125 (0.1040)	1.0123 (0.1040)	1.0384 (0.1060)	2.5935 (0.0405)	2.5987 (0.0405)
DS514Q03C	0.8229 (0.0624)	0.8228 (0.0624)	0.8349 (0.0627)	2.4033 (0.0371)	2.4077 (0.0372)
DS514Q04C	1.4247 (0.0885)	1.4239 (0.0885)	1.4478 (0.0897)	3.1050 (0.0495)	3.1008 (0.0494)
CS605Q01S	0.7084 (0.0591)	0.7085 (0.0591)	0.7066 (0.0592)	2.6121 (0.0407)	2.6099 (0.0406)
CS605Q02S	1.4711 (0.0910)	1.4693 (0.0909)	1.4660 (0.0906)	2.2272 (0.0342)	2.2428 (0.0344)
CS605Q03S	1.0544 (0.0714)	1.0539 (0.0714)	1.0635 (0.0719)	2.1355 (0.0328)	2.1315 (0.0327)
DS605Q04C	1.5606 (0.0956)	1.5612 (0.0956)	1.5615 (0.0958)	2.2107 (0.0344)	2.2066 (0.0343)
CS646Q01S	1.6558 (0.1367)	1.6541 (0.1365)	1.6904 (0.1411)	2.6925 (0.0429)	2.7045 (0.0431)
CS646Q02S	1.1099 (0.0747)	1.1098 (0.0747)	1.1081 (0.0748)	2.8149 (0.0456)	2.8327 (0.0460)
CS646Q03S	1.1524 (0.0863)	1.1520 (0.0863)	1.1571 (0.0870)	2.7632 (0.0453)	2.7748 (0.0455)
DS646Q04C	1.2392 (0.0885)	1.2393 (0.0885)	1.2284 (0.0878)	2.2151 (0.0374)	2.2349 (0.0378)
DS646Q05C	1.1760 (0.1073)	1.1766 (0.1073)	1.1554 (0.1056)	2.2537 (0.0401)	2.3056 (0.0413)
CS620Q01S	0.9229 (0.1047)	0.9222 (0.1047)	0.9326 (0.1060)	2.9307 (0.0550)	2.9417 (0.0553)
CS620Q02S	0.7000 (0.0702)	0.6999 (0.0702)	0.6999 (0.0702)	3.1221 (0.0620)	3.1355 (0.0626)
CS645Q01S	1.0283 (0.0871)	1.0286 (0.0872)	1.0214 (0.0868)	2.4280 (0.0493)	2.4383 (0.0498)
CS645Q03S	0.9556 (0.0877)	0.9554 (0.0876)	0.9569 (0.0879)	1.9649 (0.0408)	1.9695 (0.0410)
DS645Q04C	1.3153 (0.1162)	1.3148 (0.1161)	1.3171 (0.1166)	2.5370 (0.0584)	2.5864 (0.0600)
	b			γ	
CS498Q02S	0.4339 (0.0479)	0.4332 (0.0479)	0.4336 (0.0481)	0.2963 (0.0088)	0.2964 (0.0087)
CS498Q03S	0.0513 (0.0417)	0.0511 (0.0417)	0.0510 (0.0418)	−0.0035 (0.0100)	−0.0034 (0.0099)
DS498Q04C	−0.9789 (0.0585)	−0.9801 (0.0586)	−0.9774 (0.0583)	0.8081 (0.0101)	0.8073 (0.0100)
DS514Q02C	−2.8613 (0.1129)	−2.8618 (0.1130)	−2.8802 (0.1149)	0.6442 (0.0095)	0.6419 (0.0093)
DS514Q03C	0.0788 (0.0477)	0.0777 (0.0477)	0.0773 (0.0478)	0.7691 (0.0100)	0.7669 (0.0099)
DS514Q04C	−0.8505 (0.0632)	−0.8508 (0.0632)	−0.8586 (0.0638)	0.7244 (0.0084)	0.7231 (0.0083)
CS605Q01S	−0.4656 (0.0477)	−0.4659 (0.0477)	−0.4648 (0.0477)	0.5231 (0.0094)	0.5211 (0.0094)
CS605Q02S	0.1739 (0.0587)	0.1707 (0.0586)	0.1709 (0.0586)	0.1241 (0.0106)	0.1216 (0.0105)
CS605Q03S	−0.6855 (0.0548)	−0.6859 (0.0548)	−0.6849 (0.0550)	0.3177 (0.0110)	0.3141 (0.0110)
DS605Q04C	−0.6236 (0.0639)	−0.6244 (0.0639)	−0.6187 (0.0639)	0.5196 (0.0108)	0.5130 (0.0107)
CS646Q01S	−3.2734 (0.1586)	−3.2727 (0.1585)	−3.2932 (0.1621)	0.1008 (0.0094)	0.0976 (0.0093)
CS646Q02S	−0.5807 (0.0562)	−0.5813 (0.0562)	−0.5718 (0.0560)	0.2729 (0.0092)	0.2667 (0.0091)
CS646Q03S	−1.7230 (0.0783)	−1.7236 (0.0783)	−1.7124 (0.0780)	0.1989 (0.0094)	0.1923 (0.0093)
DS646Q04C	0.7145 (0.0638)	0.7136 (0.0638)	0.7307 (0.0639)	0.7706 (0.0114)	0.7423 (0.0115)
DS646Q05C	1.8926 (0.0940)	1.8919 (0.0940)	1.9070 (0.0943)	0.3988 (0.0118)	0.3687 (0.0117)
CS620Q01S	−2.5137 (0.1146)	−2.5142 (0.1146)	−2.4984 (0.1139)	−0.3169 (0.0100)	−0.3283 (0.0099)
CS620Q02S	−0.0378 (0.0580)	−0.0385 (0.0580)	−0.0193 (0.0579)	0.2735 (0.0099)	0.2559 (0.0098)
CS645Q01S	−0.2157 (0.0667)	−0.2168 (0.0667)	−0.1856 (0.0663)	0.1034 (0.0125)	0.0801 (0.0125)
CS645Q03S	−0.7810 (0.0739)	−0.7820 (0.0739)	−0.7519 (0.0731)	−0.5991 (0.0154)	−0.6226 (0.0155)
DS645Q04C	−1.0977 (0.0969)	−1.0992 (0.0969)	−1.0538 (0.0952)	0.3746 (0.0132)	0.3344 (0.0132)

TABLE 11.
The differences in estimated item parameters between different approaches based on the PISA 2018 Science items. M2–M1 means the item parameters estimated with method 2 minus that of method 1.

Item ID	a			b			α	γ
	M2-M1	M3-M1	M3-M2	M2-M1	M3-M1	M3-M2	M2-M1	M2-M1
CS498Q02S	−0.0007	0.0121	0.0127	−0.0007	−0.0003	0.0004	0.0020	0.0001
CS498Q03S	−0.0000	0.0077	0.0077	−0.0001	−0.0003	−0.0001	−0.0050	0.0001
DS498Q04C	0.0017	−0.0148	−0.0165	−0.0013	0.0015	0.0027	−0.0056	−0.0008
DS514Q02C	−0.0002	0.0259	0.0261	−0.0006	−0.0189	−0.0184	0.0052	−0.0023
DS514Q03C	−0.0000	0.0121	0.0121	−0.0011	−0.0015	−0.0005	0.0044	−0.0021
DS514Q04C	−0.0008	0.0231	0.0239	−0.0004	−0.0082	−0.0078	−0.0042	−0.0013
CS605Q01S	0.0001	−0.0018	−0.0018	−0.0003	0.0008	0.0011	−0.0022	−0.0020
CS605Q02S	−0.0018	−0.0050	−0.0033	−0.0032	−0.0029	0.0003	0.0156	−0.0025
CS605Q03S	−0.0005	0.0091	0.0096	−0.0004	0.0005	0.0009	−0.0040	−0.0035
DS605Q04C	0.0006	0.0009	0.0003	−0.0008	0.0049	0.0057	−0.0041	−0.0065
CS646Q01S	−0.0017	0.0346	0.0363	0.0007	−0.0199	−0.0205	0.0121	−0.0033
CS646Q02S	−0.0002	−0.0018	−0.0017	−0.0006	0.0088	0.0095	0.0178	−0.0063
CS646Q03S	−0.0004	0.0047	0.0051	−0.0006	0.0106	0.0112	0.0116	−0.0066
DS646Q04C	0.0001	−0.0108	−0.0109	−0.0009	0.0162	0.0171	0.0198	−0.0282
DS646Q05C	0.0006	−0.0206	−0.0212	−0.0007	0.0145	0.0151	0.0519	−0.0301
CS620Q01S	−0.0007	0.0097	0.0104	−0.0005	0.0153	0.0158	0.0110	−0.0113
CS620Q02S	−0.0001	−0.0001	0.0000	−0.0007	0.0184	0.0192	0.0134	−0.0176
CS645Q01S	0.0004	−0.0068	−0.0072	−0.0012	0.0301	0.0312	0.0102	−0.0232
CS645Q03S	−0.0002	0.0013	0.0015	−0.0010	0.0290	0.0301	0.0047	−0.0235
DS645Q04C	−0.0004	0.0018	0.0022	−0.0014	0.0439	0.0453	0.0494	−0.0402

method incorporating censoring was able to generate accurate estimates of the model parameters and the associated standard errors under all conditions, suggesting robustness of method against RT censoring-induced missingness.

One of the main interests of the current study was to evaluate the consequences of ignoring the RT censoring in the presence of NRIs. Across simulation studies, it was observed that ignoring the censoring in the joint model or the marginal model led to consistent biases in parameter estimates. The bias tended to be larger in magnitude (1) for RT model item parameters than response model item parameters, (2) for item thresholds than slopes, (3) when the correlation between speed and ability was strong, (4) when the test was short, and (5) when the proportion of NRIs was high. Jointly modeling the responses and RTs was able to mitigate the bias in response model parameter estimates to some extent, but not completely. These findings were consistent with expectations: Because the additional censoring term only involves the RT on the first unreached item for each individual, RT model parameters are directly impacted, whereas response model parameters are only indirectly affected through the correlation between speed and ability. Further, as the number of completed items increased, the relative contribution of the additional censoring term to the likelihood decreases, resulting in less severe biases from ignoring the censoring. In addition to the results reported in the present article, the performance of the proposed method and its alternatives was also evaluated under other distributions of true item parameters, for example, when the item threshold and slope were correlated, when the correlation between time intensity and item threshold was lower, and when the time discrimination parameters were larger. The results were consistent with the findings of the current study.

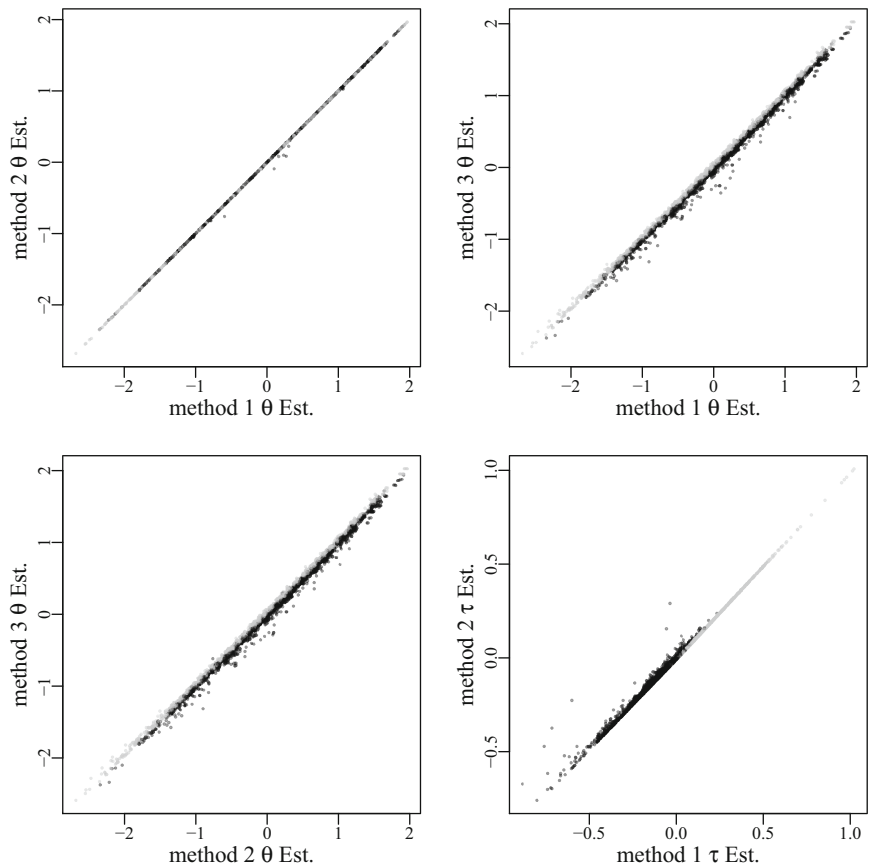

FIGURE 5.
Comparison of person parameter estimates for the PISA 2018 science data produced by different approaches. Black dots on the scatter plots represent the individuals with NRIs, and the gray dots represent individuals who have completed all part 1 items within 30 minutes.

The different approaches were further applied to the data collected from the PISA 2018 Science Test. Methods that do not explicitly model the censoring showed a consistent tendency to produce higher item threshold and lower time intensity parameter estimates. The difference in RT model item parameter estimates was larger in magnitude for the later items on the test, which had higher proportion of NRIs. Finally, comparing the two joint models (with or without censoring) on examinee latent trait estimates, ignoring the censoring was associated with lower estimates of latent abilities and higher estimates of latent speeds for individuals with at least one NRI.

4.2. Implications for Practice

When the missingness is due to reaching the time limit, for each examinee with at least one NRI, the conditional density of the observed data given the model parameters involves a censoring term. Ignoring the censoring mechanism in the likelihood function can lead to inaccurate parameter estimates and inferences. This was especially the case for the RT model parameters, which may be used for test assembly (2017) and for setting the optimal time limit of tests (van der Linden 2011). The proposed method, which explicitly includes the RT censoring term in the likelihood, has the advantage of producing unbiased and consistent estimates of item and structural parameters in the

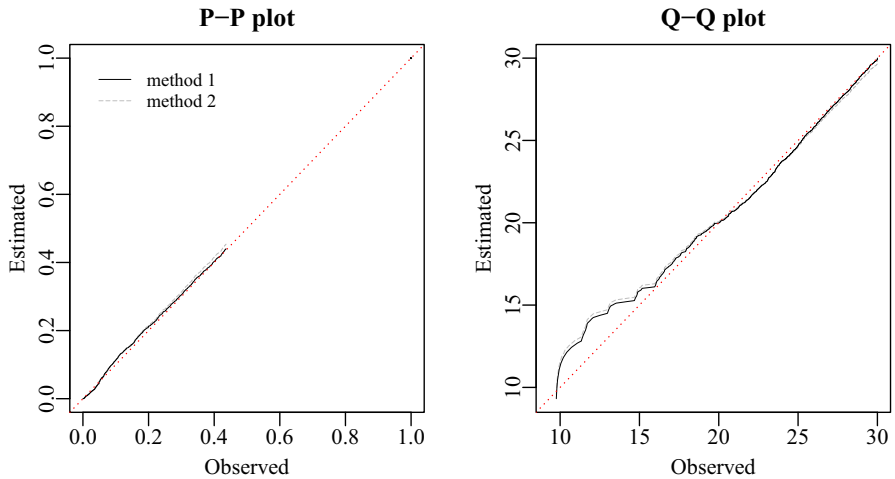

FIGURE 6.
P–P plot and Q–Q plot based on cumulative observed RTs.

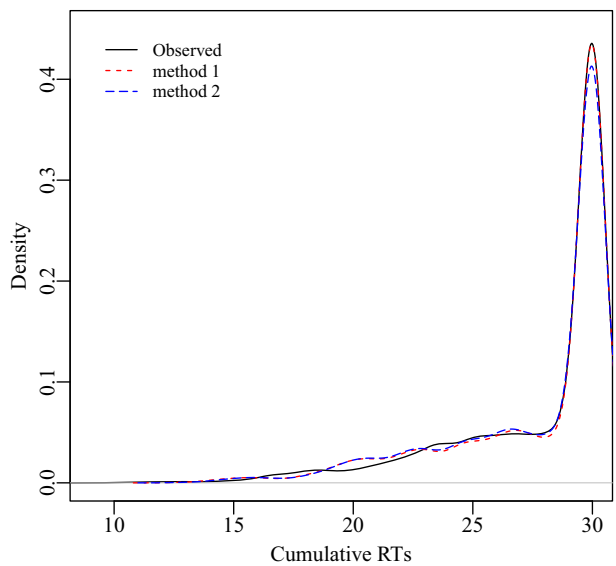

FIGURE 7.
The densities of cumulative RTs produced by the observed RTs and model-implied ones.

presence of missingness due to NRI. Because the missing mechanism is directly explained by the RT censoring, no additional parameters were introduced on top of the joint model of responses and RTs, and thus, estimation precision can be improved without increasing model complexity.

The proposed method can also improve estimates of examinees' latent abilities and latent speeds. The improvement was generally larger in magnitude for the measurement of speed than ability, as the RT censoring is directly related to the examinees' latent speeds. This is particularly useful when test developers are interested in the latent speeds of the examinees, for example, when the fluency of applying known skills is of interest to educators (e.g., 2019), or when test takers are scored or evaluated based on a composite of their latent speeds and abilities. During

a test, an examinee may balance his/her own speed and accuracy by choosing to operate at a particular speed, which implies a level of accuracy. The hierarchical model of responses and RTs (van der Linden 2007) measures the effective speed and effective ability of test takers, that is, the particular level of speed and ability they choose to operate on. Compared to focusing solely on the latent ability, assessing performance simultaneously based on effective ability and speed can provide a more complete profile of the examinees (e.g., 2018; Pohl et al. 2019). Another potential application lies in the scoring of incomplete tests in computerized adaptive testing (CAT). In CAT, subsequent items are selected on-the-fly based on provisional latent trait estimates. For examinees who do not reach the end of the test by the time limit, subsequent items have not been selected, and how to score their responses remains a practical issue. Instead of post-administration score adjustments using proportional scoring or regression-based approaches (e.g., 1996; 2001), test developers may alternatively consider scoring their performance based on a combination of speed and ability estimated from their censored responses. In such a scenario, the likelihood function with the censoring term is expected to generate more accurate speed and ability estimates.

4.3. Extensions and Further Developments

The current study has its limitations. To start with, in addition to the methods evaluated in the current study, other well-established methods for handling missing data due to NRIs, including regression-based (e.g., Rose et al. 2010; 2017) and latent variable-based (e.g., Moustaki and Knott 2000; Glas and Pimentel 2008) methods, also exist. How the proposed approach compares to these methods remains to be studied.

The proposed approach for handling NRIs can also be extended to allow for missingness due to other causes. Besides reaching the time limit, other kinds of missingness, including omissions and early quitting, have also been spotted in operational tests. Many existing studies (e.g., 2014; 2017; Lu et al. 2018; 2020) have looked into methods for handling the combination of different types of missing responses, some of which also take advantage of the RT information to explain the underlying mechanisms. In future research, the current RT censoring-induced missingness can be modeled in conjunction with other types of missing patterns, which would allow its applications in broader contexts.

Another potential extension is to permit non-stationary speed and alternative test-taking strategies within-person: The hierarchical model adopted in the current study assumes that all test takers work at a constant speed and adopt a solution strategy throughout the test. This might not be the case in practice, especially for tests with time limits. For instance, as the time limit of the examination approaches, an examinee might speed up at the cost of accuracy (e.g., 2016) or switch from a solution strategy to a rapid-guessing strategy (e.g., 2015). It is also possible that the solution processes underlying a correct response and an incorrect response differ (e.g., 2015) in a timed test. Although Remark 4 extended the current results on modeling NRI due to RT censoring to more general models for responses and RTs, the MML parameter estimator and the numerical results in the current study were based only on the hierarchical response and RT model in van der Linden (2007). Should the proposed approach be applied to other models for responses and RTs (e.g., Bolsinova et al.2017; van der Linden and Glas 2010; Wang and Hanson 2005), parameter and standard error estimators need to be derived separately based on the specific model. We leave the adaptation of the proposed approach to other joint response and RT models, as well as the evaluation of different parameter estimation methods in the presence of NRIs, to future studies.

Acknowledgments

This research was supported in part by NSF Grants DMS-2015417, SES-1826540 and IIS-1633360. Dr. Jinxin Guo (CSC No. 201906620083) and Dr. Xin Xu (CSC No. 201806620023) were sponsored by the China Scholarship Council as joint Ph.D. students.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability Statistical theories of mental test scores.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology, 71*(1), 13–38.

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.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika, 16*(2), 167–188.

Evans, F. R., & Reilly, R. R. (1972). Astudy of speededness as a source of test bias 1. *Journal of Educational Measurement, 9*(2), 123–131.

Glas, C. A., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*(6), 907–922.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., & Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement, 55*(2), 308–327.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*(1), 1–17.

Johnson, E., Allen, N. (1992). The 1990 naep technical report (no. 21-tr-20). Washington, DC: National Center for Education Statistics.

Kyllonen, P., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence, 4*(4), 14.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data (362)*. Hoboken: Wiley.

Lee, Y. H., & Ying, Z. (2015). A mixture cure-rate model for responses and response times in time-limit tests. *Psychometrika, 80*(3), 748–775.

Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media, Berlin

Little, R. J., & Rubin, D. B. (1986). *Statistical analysis with missing data*. Hoboken: John Wiley & Sons Inc.

Lu, J., Wang, C., Tao, J. (2018). Modeling nonignorable missing for not-reached items incorporating item response times. Presented at the 83rd International Meeting of the Psychometric Society, New York, NY.

Luecht, RM., Sireci, SG. (2011). A review of models for computer-based testing. Research report 2011-12. College Board.

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.

OECD. (2009). *PISA 2006 technical report*. Paris: France.

OECD. (2021). *PISA 2018 technical report*. Paris: France.

Ouircheartaigh, 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.

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.

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.

Pohl, S., Ulitzsch, E., von Davier, M. (2019). Using response times to model not-reached items due to time limits. Psychometrika1–29.

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.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in irt models. *Psychometrika, 82*(3), 795–819.

Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (irt). ETS Research Report Series, 2010(1), i–53.

Roskam, EE. (1997). Models for speed and time-limit tests. Handbook of modern item response theory (187–208). Springer.

Schleicher, A. (2019). *PISA 2018: Insights and interpretations*. OECD Publishing.

Steffen, M., Schaeffer, G. (1996). Comparison of scoring models for incomplete adaptive tests. Presentation to the Graduate Record Examinations Technical Advisory Committee for the GRE General Test.

Talento-Miller, E., Guo, F., & Han, K. T. (2013). Examining test speededness by native language. *International Journal of Testing, 13*(2), 89–104.
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.
Ulitzsch, E., von Davier, M., Pohl, S. (2019). Using response times for joint modeling of response and omission behavior. Multivariate behavioral research1–29.
Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement, 80*(3), 522–547.
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287.
van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement, 35*(3), 183–199.
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.
Veldkamp, B. P., Avetisyan, M., Weissman, A., & Fox, J. P. (2017). Stochastic programming for individualized test assembly with mixture response time models. *Computers in Human Behavior, 76*, 693–702.
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477.
Wang, S., Zhang, S., Shen, Y. (2019). A joint modeling framework of responses and response times to assess learning outcomes. Multivariate behavioral research, 1–20.
Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323–339.
Way, W. D., Gawlick, L. A., & Eignor, D. R. (2001). Scoring alternatives for incomplete computerized adaptive tests 1. ETS Research Report Series, 2001(2), i–35.
Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86–105.
Wise, S. L., Ma, L. (2012). Setting response time thresholds for a cat item pool: The normative threshold method. In annual meeting of the national council on measurement in education, Vancouver, Canada (163–183).

Manuscript Received: 2 MAR 2020
Final Version Received: 31 AUG 2021
Accepted: 10 SEP 2021
Published Online Date: 15 OCT 2021