

Item Response Theory Modeling for Examinee-selected Items with Rater Effect

Applied Psychological Measurement
2019, Vol. 43(6) 435–448
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618798667
journals.sagepub.com/home/apm



Chen-Wei Liu¹ , Xue-Lan Qiu² and Wen-Chung Wang^{2†} 

Abstract

Some large-scale testing requires examinees to select and answer a fixed number of items from given items (e.g., select one out of the three items). Usually, they are constructed-response items that are marked by human raters. In this examinee-selected item (ESI) design, some examinees may benefit more than others from choosing easier items to answer, and so the missing data induced by the design become missing not at random (MNAR). Although item response theory (IRT) models have recently been developed to account for MNAR data in the ESI design, they do not consider the rater effect; thus, their utility is seriously restricted. In this study, two methods are developed: the first one is a new IRT model to account for both MNAR data and rater severity simultaneously, and the second one adapts conditional maximum likelihood estimation and pairwise estimation methods to the ESI design with the rater effect. A series of simulations was then conducted to compare their performance with those of conventional IRT models that ignored MNAR data or rater severity. The results indicated a good parameter recovery for the new model. The conditional maximum likelihood estimation and pairwise estimation methods were applicable when the Rasch models fit the data, but the conventional IRT models yielded biased parameter estimates. An empirical example was given to illustrate these new initiatives.

Keywords

rater severity, examinee-selected items, missing not at random

Introduction

In large-scale testing, it is not uncommon to require examinees to choose and answer a fixed number of items (e.g., two) from a given set of items (e.g., four), which are referred to as examinee-selected items (ESIs). For example, several subjects in the 2016 Hong Kong Diploma of Secondary Education Examination consist of ESIs. The biology test requires examinees to choose and answer two out of the four constructed-response (CR) items. The chemistry test requires examinees to select two out of the three given sections and answer all CR items in the chosen sections. The physics, integrated science, geography, information and communication

¹The Chinese University of Hong Kong, Sha Tin, Hong Kong

²The Education University of Hong Kong, Tai Po, Hong Kong

Corresponding Author:

Chen-Wei Liu, Faculty of Education, The Chinese University of Hong Kong, Ho Tim Building, Sha Tin, New Territories, Tai Po, Hong Kong.

Email: cwliu@cuhk.edu.hk

technology, and history tests consist of ESIs as well. Other large-scale tests with ESIs include the chemistry tests in 1968 and 1969 and the history test in 2010 of the Advanced Placement Examination in the United States (Lukhele, Thissen, & Wainer, 1994; Wainer & Thissen, 1994) and the Maryland School Performance Assessment Program in the United States (Fitzpatrick & Yen, 1995), and the National Higher Education Entrance Examination in China (W. C. Wang, Jin, Qiu, & Wang, 2012). In these tests, all ESIs are in the CR format and graded by human raters.

Although several educational advantages of the ESI design have been identified, such as increasing learner autonomy, reducing test anxiety, and boosting learning (Wainer & Thissen, 1994), the measurement with the ESI design encounters two challenges—one is the problem of missing not at random (MNAR) data, and the other is the effect of rater severity on CR items. The first challenge indicates that the missing data in the ESI design (i.e., those responses to unselected items) are not ignorable in likelihood inference (Rubin, 1976). For example, more capable students (on the intended latent ability) tend to choose easier items more often than less capable students, and such choice effect makes test scores incomparable across examinees who choose different items (Lukhele et al., 1994; Wainer & Thissen, 1994). The second challenge is that the ESI design usually comprises CR items that require raters to give scores, and raters often have different degrees of severity (Linacre, 1989). Although there are attempts to address, avoid, or overcome the first challenge (Allen, Holland, & Thayer, 2005; Bradlow & Thomas, 1998; Culpepper & Balamuta, 2017; Fitzpatrick & Yen, 1995; Liu & Wang, 2017a, 2017b; Livingston, 1988; Lukhele et al., 1994; Pena, Costa, & Braga Oliveira, 2018; Powers & Bennett, 1999; W. C. Wang et al., 2012), no research has been conducted to address or overcome the second challenge for ESIs, to the best of our knowledge.

Rater errors may come from consistently giving ratings that are higher or lower than the examinees should receive (leniency/severity), overusing middle or extreme categories of a rating scale (centrality/extremity), the rater's general impression of an examinee (halo effect), or the interaction between raters (dependency; for a review, see Myford & Wolfe, 2003). Because ESIs are usually CR items that are marked by human raters, and human raters usually exhibit very different degrees of severity, it is important to consider both choice effect and rater severity to increase the feasibility of the ESI design, which is the main purpose of this study.

There are several approaches to the choice effect in the ESI design, including pattern-mixture models (Wainer & Thissen, 1994), item response theory (IRT) models with prespecified choice behaviors (Mislevy & Wu, 1996), and bifactor IRT models (W. C. Wang et al., 2012). Wainer and Thissen's pattern-mixture models assume that $\Pr(Y|Q = 0) = \Pr(Y|Q = 1)$, where Y is the response and Q is a missing datum indicator, in which $Q = 1$ if the datum is observed, and $Q = 0$ otherwise. Moreover, $\Pr(Y|Q = 0)$ is unknown unless a researcher can acquire the missing data, so the assumption of $\Pr(Y|Q = 0) = \Pr(Y|Q = 1)$ cannot be verified empirically. Mislevy and Wu's models assume that examinees' choice behaviors are known prior to data analysis, but such an assumption is unlikely to hold true in practice. In W. C. Wang et al.'s (2012) study, a latent propensity is incorporated to account for the choice effect, but the data are assumed missing completely at random (MCAR), which may not hold true in the ESI design.

In addition to these approaches, two others have been recently proposed to deal with the choice effect in the ESI design. One is the nonignorable missingness ESI model (NESIM; Liu & Wang, 2017a); the other is the conditional maximum likelihood estimation (CMLE) and pairwise estimation for the Rasch models (Liu & Wang, 2017b). The NESIM combines two IRT models: an ordinary one for substantive measures and the nominal response model (NRM; Bock, 1972) for the missingness patterns. The CMLE and pairwise estimation methods are feasible for ESIs because of the measurement property of *specific objectivity* in the family of

Rasch models, in which the estimations of item and person parameters are mutually independent. Unfortunately, both approaches fail to account for rater severity.

The purpose of this study was to advance the previous approaches to accommodate both choice effect and rater severity in the ESI design. Specifically, the authors propose (a) a new nonignorable missingness ESI rater model (NESIRM) and adapted (b) CMLE and pairwise estimation methods for ESI items to examine whether MNAR effect or rater effect exists. The former is a new IRT model for MNAR data, whereas the latter are estimation methods that can be applied to the Rasch model for MNAR data.

The authors demonstrate simulation results in the subsequent section to explicate the detriments of ignoring the MNAR effect and/or rater effect on item parameter estimators. The new methods are expected to perform well in recovering the “true” parameters when MNAR effect and/or rater effect is not ignored. Also, the new methods are expected to perform similarly to conventional methods if MNAR effect and/or rater effect are ignorable. Details are presented in the following simulation studies. In addition, the new and conventional methods were applied to empirical data to compare their difference in item estimates. Significant difference might imply that there exists MNAR effect or/and rater effect. In such situations, the new methods should yield more reliable estimates than those by the conventional methods. Details are presented in the empirical example section.

This study is organized as follows. The NESIRM and its relationship with the NESIM are introduced. Then, Rasch models for rater severity are outlined. Next that the missingness mechanism and the substantive latent trait in the ESI design could be eliminated in the CMLE and pairwise estimation methods when specific objectivity holds true is demonstrated. How to estimate the parameters of the NESIRM and implement the CMLE and pairwise estimation methods are described. Then the results of a series of simulations that were conducted to investigate the parameter recovery of the NESIRM, the effectiveness of the CMLE and pairwise estimation methods, and the consequences of ignoring choice effect and rater severity on parameter estimation are summarized, using conventional IRT models. An empirical example is provided online in appendix O to illustrate the implications and applications of the new initiatives. Finally, conclusions are drawn and suggestions for future studies are given.

The NESIRM for Choice Effect and Rater Severity

First of all, the authors introduced the necessary components of a general framework of missingness modeling for item response models. Generally speaking, an item response model has to be specified for the observed responses and a missingness model for missing data indicators for the NESIRM. Conventionally, the missing data indicator is a binary random variable used to indicate whether a response is missing (coded “0”) or observed (coded “1”). However, such coding is not appropriate to the ESI. Liu and Wang (2017a) indicated that the missing data indicators are statistically dependent due to the nature of the ESI design. Take the “choose one from two items” as an example. The resulting missing data indicators will be (1, 0) or (0, 1) for the two items. The other patterns such as (1, 1) or (0, 0) are not allowed in such design. Thus, the two missing data indicators are dependent of each other. A good choice of missing data indicators is to regard the missing patterns as nominal variables as shown in the following paragraphs.

Let Y_{com} denote complete data and consist of an observable part Y_{obs} and a missing part Y_{mis} . Y_{obs} and Y_{mis} are categorical variables in this study. Let $M_b \in (1, \dots, k, \dots, W_b)$ signify the index of the selection patterns within block b of items, and W_b represent the number of selection patterns in block b . b denotes the index of the block, and a block means a group of items that students have to select from. The random variable M_b takes on a set of possible different values. Take “choose two out of four items” as example. There can be six patterns, so $M_b = 6$. The

realized value, m_b , of M_b could be one of the six values (1, 2, 3, 4, 5, and 6). As a result, using M_b avoids the statistical dependence as mentioned previously.

Given that Y_{com} and M_b are both random variables, the joint probability of Y_{com} and M_b , $\Pr(Y_{\text{com}}, M_b)$ can be factorized as follows:

$$\begin{aligned}\Pr(Y_{\text{com}}, M_b) &= \Pr(M_b | Y_{\text{com}}) \Pr(Y_{\text{com}}) \\ &= \Pr(M_b | Y_{\text{obs}}, Y_{\text{mis}}) \Pr(Y_{\text{obs}}, Y_{\text{mis}}).\end{aligned}\quad (1)$$

By marginalizing over the unobservable Y_{mis} , the joint probability becomes,

$$\Pr(Y_{\text{obs}}, M_b) = \sum_{Y_{\text{mis}} \in S} \Pr(M_b | Y_{\text{obs}}, Y_{\text{mis}}) \Pr(Y_{\text{obs}}, Y_{\text{mis}}), \quad (2)$$

where $S \in (0, 1, \dots, C)$ and C denotes the total number of rating points minus one. By employing the parameters of interest, Equation 2 becomes

$$\Pr(Y_{\text{obs}}, M_b, \theta, \xi, \gamma, \zeta) = \sum_{Y_{\text{mis}} \in S} \Pr(M_b | Y_{\text{obs}}, Y_{\text{mis}}, \gamma, \zeta) \Pr(Y_{\text{obs}}, Y_{\text{mis}} | \theta, \xi), \quad (3)$$

where θ and γ are the target latent trait and some latent propensity (e.g., individual's tendency, which can be related to θ), respectively, ξ is the collection of all item parameters, ζ denotes the collection of all structural parameters of the missingness model. The prior distributions of item parameters (ξ and ζ) are omitted due to the absence of prior information in this paper. Let ρ denote the linear correlation between θ and γ to account for the MNAR effect, and assume that θ and γ follow a bivariate normal distribution; then Equation 3 becomes

$$\Pr(Y_{\text{obs}}, M_b, \theta, \xi, \gamma, \zeta, \rho) = \sum_{Y_{\text{mis}} \in S} \Pr(M_b | Y_{\text{obs}}, Y_{\text{mis}}, \gamma, \zeta) \Pr(Y_{\text{obs}}, Y_{\text{mis}} | \theta, \xi) \Pr(\theta, \gamma | \rho). \quad (4)$$

Furthermore, based on the local independence assumption, the M_b , Y_{obs} , and Y_{mis} are assumed stochastically independent given γ , the Y_{mis} is marginalized and Equation 4 is simplified to,

$$\begin{aligned}\Pr(Y_{\text{obs}}, M_b, \theta, \xi, \gamma, \zeta, \rho) &= \sum_{Y_{\text{mis}} \in S} \Pr(M_b | \gamma, \zeta) \Pr(Y_{\text{obs}}, Y_{\text{mis}} | \theta, \xi) \Pr(\theta, \gamma | \rho) \\ &= \Pr(M_b | \gamma, \zeta) \Pr(Y_{\text{obs}} | \theta, \xi) \Pr(\theta, \gamma | \rho),\end{aligned}\quad (5)$$

which is the NESIM (Liu & Wang, 2017a). Moreover, $\Pr(Y_{\text{obs}} | \theta, \xi)$ is assumed to follow an IRT model such as the partial credit model (PCM; Masters, 1982), whereas $\Pr(M_b | \gamma, \zeta)$ is assumed to follow the NRM. If $\rho = 0$, it leads to

$$\begin{aligned}\Pr(Y_{\text{obs}}, M_b, \theta, \xi, \gamma, \zeta, \rho) &= \Pr(M_b | \gamma, \zeta) \Pr(Y_{\text{obs}} | \theta, \xi) \Pr(\theta, \gamma | \rho = 0) \\ &= \Pr(M_b | \gamma, \zeta) \Pr(Y_{\text{obs}} | \theta, \xi) \Pr(\theta) \Pr(\gamma) \\ &\propto \Pr(Y_{\text{obs}} | \theta, \xi) \Pr(\theta),\end{aligned}\quad (6)$$

which means the missingness model can be ignored (i.e., missing data are ignorable) and this is an MCAR mechanism. The missing at random (MAR), $\Pr(M_b | \gamma, \zeta, Y_{\text{obs}})$, is not considered in the NESIM due to the local independence assumption (i.e., Y_{obs} is ignored given γ).

Notice that ρ does not convey the information about whether the choice effect in a specific block is related to θ (i.e., nonignorable). In this study, the authors relax the linear correlation assumption between θ and γ by introducing a block-specific parameter to detect the choice effect in each block.

Thus, γ is decomposed as a linear combination of θ and a new random effect, ε (e.g., individual's tendency which is not related to θ), in the multidimensional NRM (MNRM; see Equation 8) to account for the choice effect. Different from the NESIM, Equation 5 is changed as follows:

$$\Pr(Y_{\text{obs}}, M_b, \theta, \xi, \varepsilon, \zeta) = \Pr(M_b | \theta, \varepsilon, \zeta) \Pr(Y_{\text{obs}} | \theta, \xi) \Pr(\varepsilon) \Pr(\theta), \tag{7}$$

where $\Pr(\theta)$ and $\Pr(\varepsilon)$ are the distributions of θ and ε , respectively, and assumed stochastically independent to each other because a relationship of linear addition for θ and ε is assumed. Based on our experience, if theta and epsilon are assumed dependent, in addition to the linear relationship, model identification problems will occur. Moreover, $\Pr(M_b | \theta, \varepsilon, \zeta)$ for person n and choice pattern k in block b follows the MNRM):

$$\Pr(M_{bn} = k | \theta_n, \varepsilon_n, \omega_{bk}, \lambda_{bk}, \tau_{bk}) = \frac{\exp(\omega_{bk}\theta_n + \lambda_{bk}\varepsilon_n + \tau_{bk})}{\sum_{w=1}^{W_b} \exp(\omega_{bw}\theta_n + \lambda_{bw}\varepsilon_n + \tau_{bw})}, \tag{8}$$

where $\zeta \in (\omega, \lambda, \tau)$, ω_{bk} is a slope parameter for θ_n , λ_{bk} represents a slope parameter for ε_n , τ_{bk} signifies an intercept parameter, and ε_n accounts for the examinee's comprehensive propensity and is assumed to be statistically independent of θ_n . The variable ω_{bk} is the key indicator of the choice effect for pattern k in block b for θ_n to determine whether the choice effect is ignorable (i.e., whether $H_0: \omega_{bk} = 0$ is true). This information could help test designers to organize the items in the blocks to reduce the choice effect for preliminary analysis or further test development.

To account for rater severity, $\Pr(Y_{\text{obs}} | \theta, \xi)$ is assumed to follow the facets model (Linacre, 1989):

$$\Pr(y_{nis} | \theta_n, \delta_i, \eta_s) = \frac{\exp\left[y_{nis}\theta_n - \sum_{c=0}^k (\eta_s + \delta_{ic})\right]}{\sum_{w=0}^C \left\{ \exp\left[w\theta_n - \sum_{c=0}^w (\eta_s + \delta_{ic})\right] \right\}}, \tag{9}$$

where δ_{ic} is the threshold c of item i , C denotes the total number of rating points minus one ($c = 0, \dots, k, \dots, C$), $y_{nis} \in (0, 1, \dots, C)$, η_s indicates the severity of rater s , and $\delta_{i0} \equiv 0$. Combining Equations 7 to 9 creates the NESIRM.

The NESIRM is a new model for ESIs, which subsumes the old NESIM as a special case in two aspects. First, the old NESIM assumes that all raters have the same level of severity (i.e., no rater effect), whereas the NESIRM recognizes that different raters may have varying degrees of severity. Second, the ρ parameter in the NESIM indicates a universal choice effect across blocks, whereas the ω parameter in the NESIRM describes the choice effect on each block. The ρ cannot inform which block of ESIs have none, weak, or strong choice effect. In the preliminary study, the ω could help practitioners to rearrange the items in the blocks to reduce the choice effect for preliminary or further study. The NESIRM is a generalized MNAR model, which can be simplified to the NESIM when $\omega_{bk} = \lambda_{bk}\rho$ and $\sigma_\varepsilon^2 = (1 - \rho^2)\sigma_\theta^2$, given that $\gamma = \rho\theta + \varepsilon$ and $\eta_s = 0$ (see online Appendix A). By further constraining $\omega_{bk} = 0$ and $\eta_s = 0$, the NESIRM is simplified to the PCM.

Parameter Estimation for the NESIRM

The NESIRM is basically a two-dimensional IRT model because it includes two latent variables (θ and ε). For parameter estimation, a researcher can use the marginal maximum likelihood with

expectation-maximization (MML-EM) algorithm to integrate the θ and γ distributions in the likelihood function. Given a specific selection pattern vector $\mathbf{m}_n = (m_{n1}, m_{n2}, \dots, m_{nB})^T$ and a specific set of rater s , the marginal likelihood can be written as follows:

$$L_M(\xi, \zeta; \mathbf{y}, \mathbf{m}) = \prod_{s=1}^S \prod_{n=1}^N \iint \left[\prod_b^B \Pr(m_{nb} | \theta, \varepsilon, \zeta) \prod_{i \in U_{m_{nb}}} \Pr(y_{niks} | \theta, \xi) \right] f(\theta) f(\varepsilon) d\theta d\varepsilon, \quad (10)$$

where N is the number of examinees. The random variable M_{nb} takes on a set of possible different values. Take “choose two out of four items” for illustration purposes. The realized value, m_{nb} , of M_{nb} could be one of the six values (1, 2, 3, 4, 5, and 6), and the corresponding random variable of the selected item index is $U_{m_{nb}}$, which takes on a set of six possible item indexes: $u_{m_{n1}} = (1, 2)$, $u_{m_{n2}} = (1, 3)$, $u_{m_{n3}} = (1, 4)$, $u_{m_{n4}} = (2, 3)$, $u_{m_{n5}} = (2, 4)$, and $u_{m_{n6}} = (3, 4)$.

In addition to the MML-EM, a researcher can adopt the Markov chain Monte Carlo (MCMC) method, which is available for various IRT models and has been implemented in free-ware, similar to the Just Another Gibbs Sampler (JAGS; Plummer, 2003). In this study, the MCMC method was adopted via JAGS because it is easy to set up model constraints. Specifically, NESIRM is identified by constraining $\omega_{b1} = \lambda_{b1} = \tau_{b1} = 0$ for the first category (alternatively, $\sum_{w=1}^{W_b} \omega_{bw} = \sum_{w=1}^{W_b} \lambda_{bw} = \sum_{w=1}^{W_b} \tau_{bw} = 0$, Bock, 1972), $\sum_i^I \sum_{w=1}^C \delta_{ic} = 0$, and $\sum_{s=1}^S \eta_s = 0$. Thus, the mean and the variance of θ can be freely estimated. The sum-to-zero constraints are adopted to agree with the CMLE and pairwise estimation, which also use the same constraints (see the next section), so the scales are comparable between the estimation methods. In addition, based on the experience with JAGS, the authors constrain λ_{bk} to be equal across blocks (i.e., $\lambda_{bk} = \lambda_k$) and positive, that is, $\lambda_k \geq 0$ because $\lambda\varepsilon = (-\lambda)(-\varepsilon)$, to yield a stable estimation for λ . At first glance, Equation 8 seems like an over-parameterized MNRM, but actually, it is not—because θ is largely measured by item response y_{obs} , and only ω is measured by item selection m .

Apart from model constraints, the ESI design must meet the following requirements to establish a common scale (Liu & Wang, 2017a). First, at least two blocks of ESIs are needed, and there are some overlaps among examinees between blocks. Second, if there is only one block of ESIs, at least two items must be chosen (e.g., choose two out of the three items). Third, if there is only one block of two ESIs, at least one compulsory item (all examinees must answer) should be included. Fourth, if there is only one block of two ESIs but no compulsory item, at least some examinees must answer both ESIs. These requirements are not too harsh to meet in practice because multiple blocks or compulsory items are usually included in the ESI design. In addition to these requirements, the rating design should be implemented well to ensure linkage among raters for parameter estimation (Linacre, 1989).

Conditional Estimation of Rasch Models for Choice Effect and Rater Severity

In this section, the aim is not to develop a new MNAR model for choice effect and add a rater severity to the NESIRM. Instead, it is shown that, by specifying an IRT model from the Rasch family for the item responses and appropriate estimation methods, one does not need to explicitly specify the missingness model. The idea is that a researcher may find an estimator that does not involve θ and the missingness mechanism so that the item parameter estimation is independent of θ and the missingness mechanism.

Liu and Wang (2017b) showed that CMLE for the Rasch models leveraged the property of sufficient statistics for θ , so the item parameter estimation was independent of θ and the

missingness mechanism in the ESI design (Fischer, 1973; Mair & Hatzinger, 2007). In this study, CMLE is adapted to deal with both choice effect and rater severity. The details of derivation can be found online in Appendix B. Specifically, the likelihood function of the item parameters and the rater effect parameters for examinee n , given rater s , is shown in Equation 1 of Appendix B. The summation of the likelihood of all possible response patterns is also shown in Equation 2 of Appendix B. By dividing Equation 1 by Equation 2, the likelihood function, which does not involve θ and missingness model, is obtained. Thus, the choice effect can be ignored and rater effect can be estimated along with item parameters.

Pairwise Estimation of Rasch Models for Choice Effect and Rater Severity

Three variants of pairwise estimation algorithms for rater data were introduced for the Rasch models (Garner & Engelhard, 2009). However, they neither considered the missing data that were induced by an incomplete rating design nor attempted to handle ESI data. In this study, pairwise estimation algorithms are adapted to take into account both choice and rater effect in the ESI design.

In the case of a large number of items or raters, CMLE may become inefficient because of the costly, recursive computation of the elementary symmetric function (Andersen, 1970). Rasch proposed pooling all item pairs to obtain a pairwise noniterative (PWN) estimation for the item parameters (Choppin, 1968, 1985). The main purpose of the pairwise estimation is to eliminate θ by calculating the odds ratio of paired items. Choppin elaborated on the PWN and the pairwise iterative (PWI) approaches based on pairwise likelihood (Zwinderman, 1995). The pairwise eigenvector (PWE) approach proposed by Garner and Engelhard (2009) could produce item parameter estimates that were nearly identical to those from the PWN and the PWI approaches. In this study, all three approaches were adapted to the ESI design and investigated their recovery of the item and the rater parameters. The details of derivation can be found in online Appendix C. Specifically, the idea of pairwise estimations is to use distinct paired response patterns such as $(y_i = 0, y_j = 1)$ and $(y_i = 1, y_j = 0)$ on items i and j . The paired response patterns are tabulated in a paired comparison matrix C . The lower-diagonal elements, c_{ji} , and the upper-diagonal elements, c_{ij} , represent the numbers of response patterns $(y_i = 0, y_j = 1)$ and $(y_i = 1, y_j = 0)$, respectively. For the PWN, by using the ratio of the two likelihood functions of response patterns and several manipulations through Equation 5 to Equation 8 of Appendix C, item and rater estimates for binary responses are calculated directly. Similarly, estimates for polytomous items are obtained by means of Equation 11 to Equation 18. In contrast, the PWI introduced a binomial distribution to model the observed response matrix C . Thus, resolving the log-likelihood functions via Equations 19 to 23 and 24 to 28 can yield the item and rater estimates. Garner and Engelhard (2009) suggested the PWE and showed that the parameters could be derived by Equation 29. Estimating the eigenvector requires a simple recursive algorithm (Garner & Engelhard, 2009).

In summary, the CMLE and the pairwise estimation methods are not affected by the choice effect and the rater severity in the ESI context, given that the Rasch models could fit the data. On the contrary, the NESIRM is not restricted to the Rasch models and could accommodate other IRT models, such as the generalized facets model (W. C. Wang & Liu, 2007); however, specification of the MNAR mechanism is required (e.g., the flexible MNRM).

Comparison Between NESIRM and Conditional/Pairwise Estimation

The major difference between the NESIRM and the conditional/pairwise estimations is that the former must specify a missingness model for missing data patterns, but it is free to specify the IRT model for the item responses. In contrast, the conditional/pairwise estimations must specify one of the Rasch models (e.g., PCM) for the item responses, but it does not need to explicitly specify the missingness model because the missingness model can be eliminated during conditional and pairwise estimations.

Both methods are summarized in Table D of online Appendix D. For example, the CMLE, PWN, PWI, and PWE are appropriate when the (facet) Rasch models are used and the missingness model does not have to be specified. The NESIRM can include any IRT model when rater effect is involved, whereas the NESIM can also include any IRT model but it ignores the rater effect. The NESIRM/NESIM can accommodate various IRT models for item responses such as the PCM, NRM (Bock, 1972), and so on. The choice of the IRT model for item responses depends on the research interest or model fit. Although the NESIRM/NESIM must specify the missingness model, Liu and Wang (2017a) found the NRM flexible to nominal missing data indicators and robust to unknown missingness models based on simulation studies. On the contrary, conditional/pairwise estimations (not new IRT models) are somewhat restricted in practice because the Rasch models must be able to fit the item responses reasonably, although their significant advantage is that the missingness model does not need to be specified.

In summary, it is suggested that practitioners use conditional/pairwise estimations first to check whether the Rasch models can fit the item responses well. The choice effect and rater effect have been tackled in the conditional/pairwise estimations, thus one needs to check the model fit to data alone. If the Rasch models failed to characterize the data, one can resort to the NESIRM, where the IRT model for the item responses must be specified by user as well as the missingness model. The researchers' responsibility is to find an IRT model that could fit the data reasonably. For missingness model, the (M)NRM is recommended due its flexibility and robustness (Liu & Wang, 2017a).

Simulations

The motivation of the simulations was to demonstrate the detriments of ignoring the MNAR effect and/or rater effect on item parameter estimators and to show that the NESIRM and the conditional/pairwise methods could perform well in recovering "true" parameters no matter whether there exists choice effect and/or rater effect. A series of simulations is conducted to compare the CMLE, pairwise estimation, NESIRM, NESIM, and PCM in terms of the recovery of the item and the rater parameters in the ESI context.

Design and Analysis

In the ESI design, the examinees were required to choose and answer one item from a pair of items. There were four blocks (pairs) of three-point items and four raters. The thresholds δ_i for item i were generated with increasing difficulty. Specifically, δ_{i1} was generated from a uniform distribution ranging from -1.5 to 0 , whereas δ_{i2} was generated from a uniform distribution ranging from 0 to 1.5 . The average of δ across items was rescaled to zero as a model constraint. In total, 500 examinees were sampled from the standard normal distribution. Such a sample size was found sufficient to demonstrate the impact of ignoring MNAR data and the rater effect (to be shown in the "Results" section) although in practice, the sample size used in the ESI design is usually far larger than 500.

Three missingness mechanisms were considered: (a) random selection (RS), (b) linear selection (LS), and (c) nonlinear selection (NS). In the RS condition, examinees chose items randomly (each item in a pair had a 0.5 probability of being chosen). The RS served as the baseline for performance comparison. In the LS condition, the more proficient the examinee is, the higher the probability of choosing the first item in a pair, regardless of its difficulty. In total, 500 probabilities were randomly drawn from a uniform distribution ranging from 0 to 1 and sorted from low to high. Likewise, the ability levels of the 500 examinees were sorted from low to high. Then, the sorted probabilities were assigned to the sorted examinees, so that the higher the ability, the higher the probability of selecting the first item, which could be either easier or more difficult in each pair. In the NS condition, the responses were generated according to the NESIRM, in which ω_{b1} was drawn from a uniform distribution between -2 and 2 , λ_{b1} was set at 1 for simplicity (it was less interesting than other parameters), and τ_{b1} was set at 0, so the relationship between item selection M and θ was nonlinear.

The severity levels of the four raters were set as $\eta = (0, -1, 0, 1)$ for simplicity, and the sum of η was set at 0 as a model constraint. Also $\eta = (0, 0, 0, 0)$ is set to indicate no rater effect. An incomplete and spiral-like rating designs were adopted (DeCarlo, 2010; Engelhard, 1997). In the incomplete rating design, each examinee was judged by two raters, and each rater judged a subset of examinees on all eight items, with overlaps between raters. Specifically, the first two raters (severity = 0, -1) judged examinees 1 to 275, whereas the last two raters (severity = 0, 1) judged examinees 226 to 500, and the overlaps between the first and the last two raters comprised 50 examinees. In the spiral-like rating design, each rater judged four of the eight items, as shown in Table E of online Appendix E. Specifically, Item 1 was marked by Raters 1 and 2, Item 2 by Raters 2 and 3, and so on until Item 8 by Raters 1 and 4. The authors did not consider a complete rating design in this study because it is seldom adopted in large-scale testing.

There were altogether 12 conditions, including two rating designs \times two rater effects \times three missingness mechanisms. In total, 70 replications were conducted under each condition, which appeared sufficient based on our preliminary studies. The item and the rater parameters were fixed across replications, whereas the person parameters were randomly drawn in each replication. For the assessment of parameter recovery, the bias and the root mean square error (RMSE) of estimator $\hat{\xi}$ were computed as $R^{-1} \sum_{r=1}^R (\hat{\xi} - \xi)$ and $\sqrt{R^{-1} \sum_{r=1}^R (\hat{\xi} - \xi)^2}$, respectively, where $R = 70$, and ξ is the true parameter.

The CMLE method implemented in the **eRm** package of the *R* program (Mair & Hatzinger, 2007) was used in this study. The PWN, PWI, and PWE algorithms were implemented on the *R* program. The NESIRM, NESIM, and PCM were fit by using the freeware JAGS, in which the burn-in period covered 4,000 iterations, and the samples included 6,000 iterations, which were found sufficient, based on our prior experiments and the following results (see the next section). The mean threshold and the mean rater severity were set at 0 for model identification. online Appendix F provides a template of the JAGS syntax for the NESIRM. Note that NESIM can be obtained by constraining $\eta = 0$ of the NESIRM and can obtain the PCM by constraining $\omega = 0$ and $\eta = 0$ of the NESIRM, so model comparison is readily available.

It was anticipated that all methods would perform satisfactorily when there was no choice effect and no rater effect. The CMLE, the pairwise estimation, and the NESIRM would yield practically unbiased estimates in all conditions. The NESIM would produce biased estimates when the rater severity existed but was ignored. The PCM would suffer more serious bias when both choice effect and rater severity existed but were ignored.

Results

This section summarizes the bias and the RMSE for the parameter estimates yielded by the seven methods (CMLE, PWN, PWI, PWE, NESIRM, NESIM, and PCM) under the incomplete and spiral-like rating designs.

Incomplete Rating Design

The bias and the RMSE of $\hat{\delta}$ and $\hat{\eta}$ for the seven methods are shown in Figures 1 and Figure G of online Appendix G, respectively. When there was no choice effect and no rater effect (Figures 1a and Ga for bias and RMSE, respectively), all methods yielded good parameter recovery. When there was a choice effect due to LS but no rater effect (Figures 1b and Gb for bias and RMSE, respectively), only the PCM yielded biased item parameter estimates and large RMSE values. Specifically, the PCM underestimated the threshold parameters for the first item of each pair by approximately 0.25 logits and overestimated those for the second item of each pair by approximately 0.25 logits. The reason was that high-ability examinees tended to select the first item, and low-ability examinees were inclined to choose the second item. When there was a choice effect due to NS but no rater effect (Figures 1c and Gc for bias and RMSE, respectively), the bias patterns yielded by the PCM depended on the choice-effect indicator ω . When there was a rater effect but no choice effect (Figures 1d and Gd for bias and RMSE, respectively), both the PCM and the NESIM yielded bias up to approximately 0.35 logits in the absolute value and RMSE larger than that obtained from the other methods by approximately 0.1. When there was a choice effect (due to LS) and a rater effect (Figures 1e and Ge for bias and RMSE, respectively), the PCM yielded large bias and RMSE for some item parameters, and the NESIM yielded large bias and RMSE for all item parameters. When there was a choice effect (due to NS) and a rater effect (Figures 1f and Gf for bias and RMSE, respectively), the results were similar to those when there was a choice effect (due to LS) and a rater effect. Across all conditions, the CMLE, PWN, PWI, PWE, and NESIRM methods yielded good recovery for the item and the rater parameters.

Spiral-Like Rating Design

The bias and the RMSE of $\hat{\delta}$ and $\hat{\eta}$ for the methods are shown in online Appendices H and I, respectively. Similar to those observed in the incomplete rating design, the PCM yielded serious bias and RMSE when there was a choice effect (Appendices Hb, Hc, Ib, and Ic). When both choice and rater effects were present, the PCM and the NESIM yielded serious bias and RMSE (Appendices Hd, He, Hf, Id, Ie, and If). The CMLE, PWN, PWI, PWE, and NESIRM methods yielded good recovery for the item and the rater parameters in all conditions. Compared to the incomplete rating design, the spiral-like rating design tended to generate larger bias and RMSE because it had sparser linkage between raters.

Appendices J and K online show the bias and the RMSE for $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\omega}$, $\hat{\lambda}$, and $\hat{\tau}$ under the incomplete rating design and the spiral-like rating design, respectively, under the NS condition. The bias and the RMSE for $\hat{\sigma}^2$ and $\hat{\omega}$ were very large for the NESIM and the PCM methods when there was a rater effect under the spiral-like rating design (Appendix Kb and Kd), which might be due to the multiplier effect of ignoring rater severity and the sparse rater linkage (Liu & Wang, 2017a).

To summarize, the CMLE, pairwise estimation, and NESIRM methods are very robust against different missingness mechanisms (i.e., RS, LS, and NS). On the contrary, the PCM and the NESIM methods always yield biased estimates when the choice effect and/or the rater

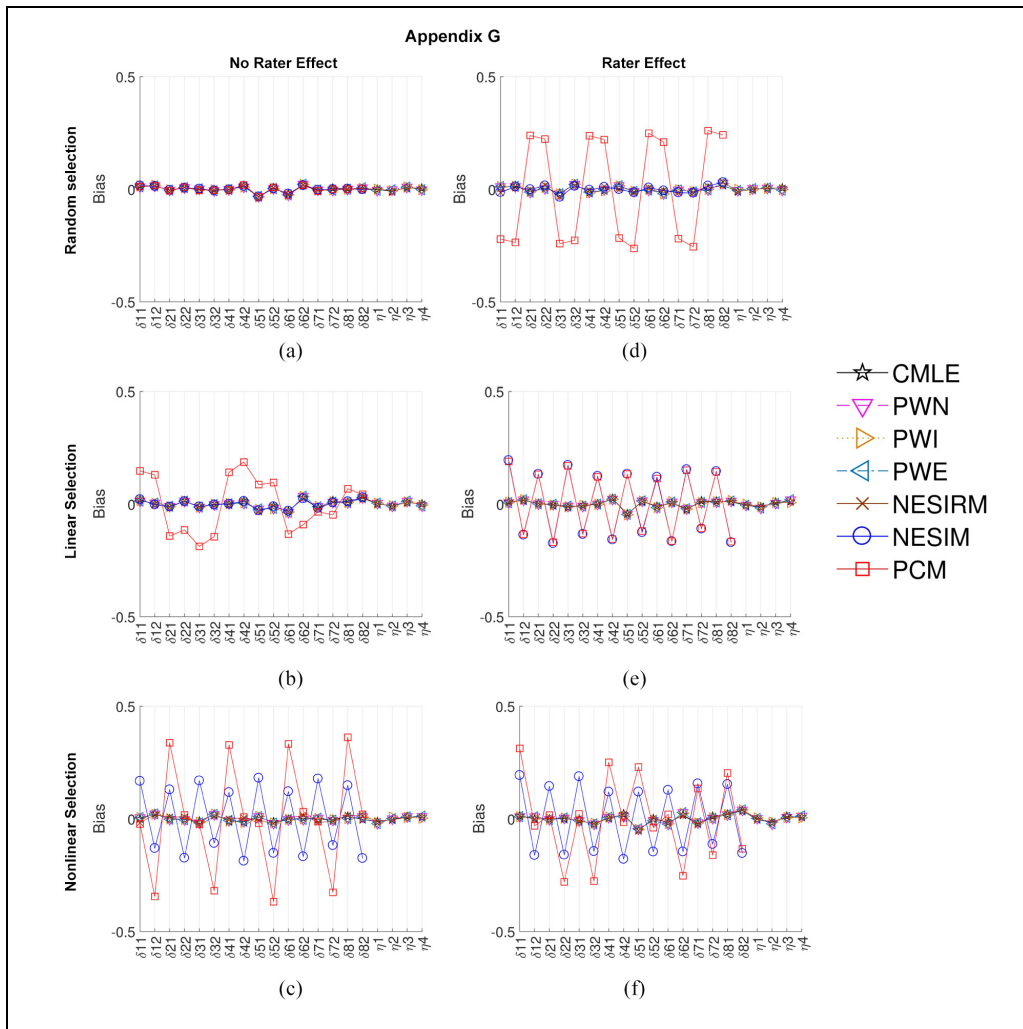


Figure 1. Bias in item estimators $\hat{\delta}$ and rater estimators $\hat{\eta}$ with the incomplete rating design. Note. (a) and (d) are for the random selection condition; (b) and (e) are for the linear selection condition; (c) and (f) are for the nonlinear selection condition; (a), (b), and (c) are for no rater effect; (d), (e), and (f) are for rater effect; $\hat{\delta}_{11}$ and $\hat{\delta}_{12}$ denote the first and second threshold of Item 1, and so on for the others; $\hat{\eta}_1$ denotes the severity of Rater 1, and so on for the others.

effect are/is present. The CMLE and the pairwise methods are useful for the ESI design with CR items when the Rasch models can reasonably fit the data. If the Rasch models do not fit the data, the NESIRM should be adopted, in which the IRT model for the observed data can be more general than the Rasch models, such as the generalized facets model. Generally, using the MNRM to account for missingness mechanism is sufficiently robust.

Other Simulation Conditions

The above simulation studies did not address the following conditions: (a) random assignment of examinees to raters, (b) small sample size, and (c) the recovery of person parameters. Thus,

three follow-up simulations are added. The goal is to examine whether the bias patterns would be affected under the three conditions, compared with previous simulation results. Details of designs and results were described in online Appendices L, M, and N.

In summary, the bias patterns were similar to previous simulation results irrespective of random assignment of examinees to raters or small sample size, whereas the person parameters can be well recovered by maximum likelihood estimation.

Conclusion and Discussion

Currently, the ESI design is rarely used in Western countries because the missing data are usually MNAR, which invalidates the use of common IRT models (X. B. Wang, Wainer, & Thissen, 1995). This conundrum inevitably overrides the practical advantages of the ESI design. Although some Asian countries still adopt the ESI design in high-stake and large-scale tests, number correct scores are usually used for score reports, completely ignoring both choice and rater effects. Despite the development of a few IRT models and approaches to tackle the choice effect in the ESI design (Liu & Wang, 2017a, 2017b), the rater effect is not considered in the literature. Because ESIs are usually in the CR format and thus graded by human raters, approaches that do not consider the rater effect become inapplicable.

In this study, the authors developed the NESIRM and adapted the CMLE and three pairwise estimation methods to account for both choice and rater effects in the ESI design. With this approach, the person parameters become comparable among examinees who choose different ESIs and whose answers are graded by different raters. Simulation studies confirm the advantages. The CMLE and the three pairwise methods require a good fit of the Rasch models, whereas the NESIRM is more flexible to accommodate other IRT models for observed data. Conventional approaches, such as the PCM and the NESIM, fail to consider the choice effect and/or the rater effect and yield biased estimates for the item parameters. With these approaches, the person parameters are not comparable among examinees who choose different ESIs and/or whose answers are graded by different raters.

The empirical example in online Appendix O illustrates a way to adopt the NESIRM, CMLE, and pairwise estimation methods to analyze ESI data with the rater effect. Taking those parameter estimates obtained from the NESIRM-C as representing the gold standard, the authors find that those obtained from the NESIRM, CMLE, and pairwise estimation methods are almost identical to the gold standard, justifying their applicability to the ESI design with the rater effect. In contrast, ignoring the choice effect and/or the rater effect by adopting the NESIM or the Rasch model results in misleading parameter estimates.

In practice, the missingness mechanisms in the ESI design may be more complex than those manipulated in this and past studies. Future studies should investigate how the NESIRM, CMLE, and pairwise estimation methods will perform under various choice effects. There are rater effects other than severity, such as inconsistency, centrality/extremity, halo, and rater dependency. The hierarchical rater model with signal detection theory rater components for the ESI design is another interesting route to explore (Patterson, 2013). Several IRT models have tackled these rater effects (DeCarlo, Kim, & Johnson, 2011; Patz, Junker, Johnson, & Mariano, 2002; W. C. Wang, Su, & Qiu, 2014; W. C. Wang & Wilson, 2005). It is important for future studies to incorporate these models into the NESIRM and evaluate its performance.

Acknowledgments

The first author would like to thank Prof. Wen-Chung Wang, and Dr. Xue-Lan Qiu for their valuable revisions and data collection for this article. This work is the first author's last collaboration with Prof. Wang. Goodbye, we will miss you


Declaration of Conflicting Interests


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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Grants Council of the Hong Kong SAR under GRF Project No. 18613716.

ORCID iDs

Chen-Wei Liu  <https://orcid.org/0000-0003-2826-465X>

†Wen-Chung Wang  <https://orcid.org/0000-0001-6022-1567>

Supplemental Material

Supplemental material is available for this article online.

References

- Allen, N. L., Holland, P. W., & Thayer, D. T. (2005). Measuring the benefits of examinee-selected questions. *Journal of Educational Measurement, 42*, 27-51.
- Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B, 32*, 283-301.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.
- Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics, 23*, 236-243.
- Choppin, B. (1968). Item bank using sample-free calibration. *Nature, 219*, 870-872.
- Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education, 9*, 29-42.
- Culpepper, S. A., & Balamuta, J. J. (2017). A hierarchical model for accuracy and choice on standardized tests. *Psychometrika, 82*, 820-845.
- DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report RR-10-08), Princeton, NJ: Educational Testing Service.
- DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333-356.
- Engelhard, G. Jr., (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*, 19-33.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.
- Fitzpatrick, A. R., & Yen, W. M. (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement, 32*, 243-259.

- Garner, M., & Engelhard, G. (2009). Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments. *Journal of Applied Measurement, 10*, 30-41.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.
- Liu, C.-W., & Wang, W.-C. (2017a). Non-ignorable missingness item response theory models for choice effects in examinee-selected items. *British Journal of Mathematical and Statistical Psychology, 70*, 499-524.
- Liu, C.-W., & Wang, W.-C. (2017b). Parameter estimation in Rasch models for examinee-selected items. *Journal of Educational Measurement, 54*, 518-549.
- Livingston, S. A. (1988). *Adjusting scores on examinations offering a choice of essay questions* (ETS Research Report RR-88-64), Princeton, NJ: Educational Testing Service.
- Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234-250.
- Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.
- Patterson, B. F. (2013). *Examining the impact of examinee-selected constructed response items in the context of a hierarchical rater signal detection model*. New York, NY: Academic Commons, Columbia University.
- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341-384.
- Pena, C. S., Costa, M. A., & Braga Oliveira, R. P. (2018). A new item response theory model to adjust data allowing examinee choice. *PLOS ONE, 13*(2), e0191600. doi:10.1371/journal.pone.0191600.
- Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.
- Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education, 12*, 257-279.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.
- Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159-195.
- Wang, W.-C., Jin, K.-Y., Qiu, X.-L., & Wang, L. (2012). Item response models for examinee-selected items. *Journal of Educational Measurement, 49*, 419-445.
- Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67*, 583-605.
- Wang, W.-C., Su, C. M., & Qiu, X. L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement, 51*, 260-280.
- Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*, 296-318.
- Wang, X.-B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8*, 211-225.
- Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19*, 369-375.