METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

A General Framework for Modeling Missing Data Due to Item Selection With Item Response Theory

Paul A. Jewsbury [1], Ru Lu [1], Peter W. van Rijn [2]

[1] *ETS Research Institute, ETS, Princeton, NJ, USA.* [2] *ETS Global, Amsterdam, The Netherlands.*

Methodology, 2024, Vol. 20(3), 218–237, https://doi.org/10.5964/meth.14823

Received: 2024-06-07 • **Accepted:** 2024-09-11 • **Published (VoR):** 2024-09-30

Handling Editor: Jone Aliri, University of the Basque Country, Leioa, Spain

Corresponding Author: Paul A. Jewsbury, ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541 USA, E-mail: pjewsbury@ets.org

Abstract

In education testing, the items that examinees receive may be selected for a variety of reasons, resulting in missing data for items that were not selected. Item selection is *internal* when based on prior performance on the test, such as in adaptive testing designs or for branching items. Item selection is *external* when based on an auxiliary variable collected independently to performance on the test, such as education level in a targeting testing design or geographical location in a nonequivalent anchor test equating design. This paper describes the implications of this distinction for Item Response Theory (IRT) estimation, drawing upon missing-data theory (e.g., Mislevy & Sheehan, 1989, https://doi.org/10.1007/BF02296402; Rubin, 1976, https://doi.org/10.1093/biomet/63.3.581), and selection theory (Meredith, 1993, https://doi.org/10.1007/BF02294825). Through mathematical analyses and simulations, we demonstrate that this internal versus external item selection framework provides a general guide in applying missing-data and selection theory to choose a valid analysis model for datasets with missing data.

Keywords

item parameter estimation, item response theory, missing data, adaptive testing, targeted testing, equating

Many problems in education testing involve item selection or routing of examinees to alternate test forms. For example, adaptive testing designs select items based on information *internal* to the test, such as an interim estimate of proficiency based on performance on earlier stages of the test (Wainer & Dorans, 2000). Another example of internal item selection is the branching technology-enhanced item type where examinees are routed to items based on responses to prior items (Wainer & Kiely, 1987). Like adaptive designs, targeted testing designs also select items to better match the difficulty of the test to the proficiency of the examinee, but use auxiliary information *external* to the test, such as education level (e.g., Yamamoto *et al.*, 2018). Another example of external item selection is assigning different test forms to examinees at different geographical locations or at different times, such as with a nonequivalent anchor test (NEAT; Holland, 2007) design. Notably, missing data occurs in all of these designs as not all items are administered to all examinees. We introduce the internal versus external distinction and show how classifying designs with respect to this distinction provides an intuitive and general framework to assist researchers in applying statistical research (e.g., Glas, 1988; Meredith, 1993; Mislevy & Sheehan, 1989; Mislevy & Wu, 1996) to choose valid methods for datasets with missing data.

Our focus is on item parameter estimation using likelihood inference (Rubin, 1976). In this context, two suitable methods are conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation. Zwitser and Maris (2015) demonstrated that CML can be used for estimating item parameters of the Rasch model for multistage testing, a type of adaptive testing design, with some adjustment of the conditional likelihood based on the routing decisions. For targeted testing designs, regular CML works well (Eggen & Verhelst, 2011). We focus on MML estimation, because MML estimation can deal with a wider range of item response theory (IRT) models, including multidimensional models. Our conclusions also apply to Bayesian inference (Rubin, 1976).

Previous research found biased or implausible item parameter estimates when using certain methods for real and simulated data generated by adaptive testing, targeted testing, or a mix of both (e.g., Lu et al., 2017, 2018; Wu & Lu, 2017; Wu & Xi, 2017; Yamamoto, Shin, & Khorramdel, 2018, Yamamoto, Khorramdel, & Shin, 2018). While various other explanations were provided, this bias may be explained in terms of statistical theory as resulting from violated model estimation assumptions (e.g., Eggen & Verhelst, 2011; Glas, 1988; Meredith, 1993; Mislevy & Sheehan, 1989; Mislevy & Wu, 1996). Consequently, a general framework to classify missing data may be useful in selecting a valid method in accordance with statistical theory to obtain valid item parameter estimates in datasets with missingness.

In this paper, we first describe the item response theory model to define our notation. The concept of internal versus external item selection is introduced and formally defined in relation to conditional (in)dependence assumptions. In the Appendix, we show that this distinction allows for simple proofs of valid and invalid item parameter estimation under MML. In the Missing-at-Random Assumption and Equality Assumption sections in the main text, we show how the distinction can be used to apply missing data and selection theories from the literature. In the Analysis model section, we summarize the principles derived from the literature in choosing a valid estimation model for datasets with missing data. Finally, we illustrate the principles with a simulation and end in a discussion.

Item Response Theory

Let y_j be a possible value of the response Y_j to item j. Let θ denote an unobserved or latent continuous variable that represents examinee proficiency. According to IRT models, the probability of a response is a function of θ and the item parameters $\boldsymbol{\beta}_j$ that characterize this relationship. That is,

$$p(Y_j = y_j | \theta, \boldsymbol{\beta}_j) \equiv p(y_j | \theta, \boldsymbol{\beta}_j)\,.$$

The functional form of this relationship depends on the specific IRT model for each item. For example, the two-parameter logistic model (2PLM) for an item j with possible responses $Y_j = 1$ (correct) or $Y_j = 0$ (incorrect) has the form,

$$p(y_j | \theta, \beta_j) = \frac{\exp(y_j a_j(\theta - b_j))}{1 + \exp(a_j(\theta - b_j))},$$

where $\boldsymbol{\beta}_j = \{a_j, b_j\}$ for the discrimination parameter a_j and difficulty parameter b_j.

IRT models assume *local independence*, a type of conditional independence between the item responses. Specifically,

$$p(\mathbf{y} | \theta, \boldsymbol{\beta}) = \prod_{j=1}^{J} p(\mathbf{y}_j | \theta, \boldsymbol{\beta}_j),$$

where $\mathcal{J}$ is the number of items, $\mathbf{y} = \{y_1, \ldots, y_J\}$ is a vector of item responses for all $\mathcal{J}$ items, and $\boldsymbol{\beta}$ represents the item parameters for all items.

Internal Versus External Item Selection

Both internal and external item selection involves routing students to alternate test forms, resulting in missing data for some items. Let the variable G denote the routing decision for the test taker. For example, in a simple multistage or targeted testing design, G may have values g that denote routing to easier items, denote routing to items of intermediate difficulty, or denote routing to hard items. In an item-level computerized adaptive test design, G may take a large number of values as there are a large number of possible test forms an examinee may receive. In a typical equating problem where different test forms are provided to examinees in different locations or at different times, G may take a small number of values, and the different test forms may or may not be equivalent in terms of difficulty.

Because internal item selection is based directly on the item responses or a function of the item responses such as an interim proficiency estimate, internal item selection is characterized by the conditional independence,

$$p_{\text{int}}(g \mid \theta, \mathbf{y}) = p_{\text{int}}(g \mid \mathbf{y}), \tag{1}$$

and by the conditional dependence,

$$p_{\text{int}}(g \mid \theta, \mathbf{y}) \neq p_{\text{int}}(g \mid \theta)\,. \tag{2}$$

In contrast, because external item selection is based on some external variable related to θ but collected independently to the assessment, external item selection is characterized by the conditional dependence,

$$p_{\text{ext}}(g \mid \theta, \mathbf{y}) \neq p_{\text{ext}}(g \mid \mathbf{y}), \tag{3}$$

and because of local independence in IRT models, by the conditional independence,

$$p_{\text{ext}}(g \mid \theta, \mathbf{y}) = p_{\text{ext}}(g \mid \theta)\,. \tag{4}$$

In internal item selection, the routing decision is characterized by being independent of latent proficiency conditional on the item responses (Equation 1), but not independent of the item responses conditional on latent proficiency (Equation 2). Conversely, in external item selection, the routing decision is characterized by the routing decision being independent of the item responses conditional on latent proficiency (Equation 4), but not independent of latent proficiency conditional on the item responses (Equation 3).

As described below, classification of item selection as either internal or external can be used to apply missing data and selection theories to model datasets with missing data. A formal treatment is provided in the Appendix.

Missing-at-Random Assumption

Internal and external item selection data is incomplete, because examinees do not receive all of the items. The present paper focuses on two of the most common methods to estimate IRT models, marginal maximum likelihood (MML; Bock & Aitkin, 1981), and Bayesian inference. These methods are valid for incomplete data when the missing data satisfies Rubin's (1976) ignorability principle.

The ignorability principle is conveniently discussed with respect to the missing-at-random (MAR) assumption and the assumption that the parameters that characterize the missing-data mechanism (ϕ) and the parameters of the model (e.g., $\boldsymbol{\beta}$) are *distinct* (D). When both assumptions are satisfied, the missing data is ignorable with MML and Bayesian inference (Rubin, 1976). In education testing, the D assumption is usually clearly satisfied and will not be discussed in detail, but the MAR assumption requires more consideration. The MAR assumption is,

$$p(\mathbf{M} = \mathbf{m} \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}, \mathbf{Y}_{\text{mis}} = \mathbf{y}_{\text{mis}}) \equiv p(\mathbf{m} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) = p(\mathbf{m} \mid \mathbf{y}_{\text{obs}}),$$

where $\mathbf{M}$ is a vector variable with associated values $\mathbf{m}$, comprising a missing-data indicator for every item. Every missing-data indicator, or element of $\mathbf{m}$, is 1 if the item response for that corresponding item is missing, and 0 if the item response is observed. $\mathbf{Y}_{\text{obs}}$ is a vector variable with associated values $\mathbf{y}_{\text{obs}}$, comprising all of the responses that were observed. $\mathbf{Y}_{\text{mis}}$ is a vector variable with associated values $\mathbf{y}_{\text{mis}}$, comprising the responses that were not observed. In the context of internal and external item selection, $\mathbf{y}_{\text{mis}}$ are the responses that would have been observed for the items the examinee was not presented with.

Mislevy and Sheehan (1989) indicated that if information used to select items is also related to examinee characteristics, as with internal and external item selection, then that information must be used when estimating the item parameters to ensure that the missing data is ignorable. In other words, when information is used in the selection of items, that information can be said to be related to $\mathbf{m}$. When that information is also related to the missing responses, $\mathbf{y}_{\text{mis}}$, a relationship between $\mathbf{m}$ and $\mathbf{y}_{\text{mis}}$ is implied. To ensure that MAR is satisfied, such information involved in the selection of items must be used when estimating the item parameters.

In internal item selection designs, the information used to select items is independent of examinee characteristics (i.e., θ) conditional on the information included in the estimation (see Equation 1). For this reason, MAR is automatically satisfied with internal item selection data, provided that all of the routing item responses are included in the estimation. For example, Glas (1988) and Mislevy and Wu (1996) showed that for unidimensional adaptive tests, a type of internal item selection design, the missing data resulting from item selection satisfies Rubin (1976) ignorability principle when IRT is used. Jewsbury and van Rijn (2020) showed that when item selection is based on an interim proficiency estimate calculated from responses to items from multiple domains, another example of internal item selection, the resulting missing data satisfies the ignorability principle when multidimensional IRT is used.

In external item selection designs, the information used to select items is dependent on examinee characteristics (i.e., θ) conditional on the information included in the estimation (see Equation 3). Information related to θ is used in item selection and is not already included within $\mathbf{y}_{\text{obs}}$, so MAR is not automatically satisfied in the context of IRT. An appropriate approach must be taken to ensure that the external information is taken into account in the item parameter estimation. For example, as discussed in the next section, a grouping variable may be used together with the latent variable (Mislevy & Wu, 1996).

Equality Assumption

In the multigroup IRT model, model identification is often achieved by assuming some of the items presented to all of the groups functions in the same way for each of the groups (Bock & Moustaki, 2007). Specifically,

$$p(\mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}} | \theta, Z = z) \equiv p(\mathbf{y}_{\text{obs}} | \theta, z) = p(\mathbf{y}_{\text{obs}} | \theta),$$

where Z is the grouping variable with values z. This is the *equality assumption*. Note that Z is the grouping variable that defines the groups in the multigroup IRT model (a property of the model) whereas G is the grouping variable that defines the routing groups (a property of the data). Therefore, $Z = G$ is only true if the routing groups are used as the grouping variable in the model.

Violation of the equality assumptions is an example of *differential item functioning* or *measurement non-invariance*. If the equality constraints are violated for an item, the item can be said to not operate the same way for all z.

The use of routing-group membership, g, as a grouping variable with θ is motivated to ensure that the missing mechanism is appropriately accounted for (Mislevy & Sheehan, 1989). Selection theory (Meredith, 1964, 1993) provides a useful framework to understand the equality assumptions with respect to internal and external item selection designs when g is used as a grouping variable.

Selection theory states that the equality assumption is violated when examinees are assigned values of z (g), if the selection is related to the item responses conditional on θ. Essentially, selection must be independent to the item response conditional on θ to ensure that the effects of selection are expressed through θ and not through differential relationships between the item responses and θ (Meredith & Tersei, 2006). When the effects of selection are expressed through differential relationships between the item responses and θ, measurement non-invariance is present. Selection theory can be understood as an extension of the classic Pearson-Lawley selection formula (Lawley, 1943).

Selection is related to the item responses conditional on θ when the item responses have measurement non-invariance in the usual sense, but also when selection is simply directly based on the item responses. For example, Muthén (1989) described how selecting a grouping variable directly based on the item responses violate the equality assumptions in the context of factor analysis, and the same general principles apply to IRT.

In internal item selection designs, selection is based directly on the item responses (see Equation 3). Selection theory means that the use of g as a grouping variable with internal item selection data produces a model with equality assumptions that are not satisfied. As g is selected directly on the item responses, using g as a grouping variable will violate the equality assumptions. For example, examinees routed to harder items are more likely to have a correct response on any given routing item than examinees routed to easier items, even when matched on θ, as routing is a direct function of the routing item responses.

In external item selection designs, selection is only indirectly related to the item responses, through θ (see Equation 4). Selection theory means that the use of g as a grouping variable with external item selection data results in equality assumptions that are satisfied. Although g is related to the item responses, local independence means that g is independent to the item responses conditional on θ. The exception to this rule is when items already have measurement non-invariance with respect to g.

Analysis Model

Consideration of the missing at random and equality assumptions shows that distinct analysis models are required for internal and external item selection data. With internal item selection, the missing at random assumption is automatically satisfied. The missing data and by extension routing group membership, g, can be ignored. Furthermore, because routing group membership is directly based on item responses, attempting to account for this information by using g as a grouping variable in a multigroup IRT will introduce erroneous assumptions. Therefore, routing-group membership can and should be ignored with internal item selection, and a single-group IRT model can be fit to internal item selection data.

In contrast, the missing at random assumption is not automatically satisfied with external item selection. The missing data, and by extension, routing group membership, cannot be ignored. Unlike internal item selection, routing group membership is not directly based on item responses, so using g as a grouping variable in a multigroup IRT will not necessarily introduce erroneous assumptions unless in the presence of measurement non-invariance. Therefore, routing group membership should be accounted for in external item selection designs, such as with a multigroup IRT model (Bock & Zimowski, 1997).

Aside from multigroup IRT models, a range of methods under the umbrella of test equating with anchor items could be used with external item selection (e.g., Kolen & Brennan, 2014; von Davier & von Davier, 2007). While evaluating the assumptions for all of these methods is beyond the scope of this paper, careful consideration should be taken before applying these methods to an internal item selection design, due to the critical differences between internal and external item selection as described in this paper.

The theoretical differences between internal and external item selection data are summarized in Table 1. To illustrate the importance of fitting the appropriate model for internal versus external item selection and to show that item parameter estimation is unbiased when the appropriate model is fit, a simulation study was conducted and described in the following section.

Table 1

Summary of Internal Versus External Item Selection

Internal item selection	**External item selection**
Characteristic conditional independency	Characteristic conditional dependency
$p_{\text{int}}(g \mid \theta, \mathrm{y}) = p_{\text{int}}(g \mid \mathrm{y})$	$p_{\text{ext}}(g \mid \theta, \mathrm{y}) \neq p_{\text{ext}}(g \mid \mathrm{y})$
MAR assumption satisfied	MAR assumption violated
$p_{\text{int}}(\boldsymbol{m} \mid \mathrm{y}_{\text{obs}}, \mathrm{y}_{\text{mis}}) = p_{\text{int}}(\boldsymbol{m} \mid \mathrm{y}_{\text{obs}})$	$p_{\text{ext}}(\boldsymbol{m} \mid \mathrm{y}_{\text{obs}}, \mathrm{y}_{\text{mis}}) \neq p_{\text{ext}}(\boldsymbol{m} \mid \mathrm{y}_{\text{obs}})$
Characteristic conditional dependency	Characteristic conditional independency
$p_{\text{int}}(g \mid \theta, \mathrm{y}) \neq p_{\text{int}}(g \mid \theta)$	$p_{\text{ext}}(g \mid \theta, \mathrm{y}) = p_{\text{ext}}(g \mid \theta)$
Equality assumption violated	Equality assumption satisfied
$p_{\text{int}}(\mathrm{y} \mid \theta, g) \neq p_{\text{int}}(\mathrm{y} \mid \theta)$	$p_{\text{ext}}(\mathrm{y} \mid \theta, g) = p_{\text{ext}}(\mathrm{y} \mid \theta)$
Valid MML analysis model	Valid MML analysis model
One group IRT model, ignoring g	Multigroup IRT model, using g

Note. MAR = missing at random, int = internal item selection, ext = external item selection.

Simulations

The simulation illustrates the principles of internal item selection with a multistage testing (MST) design, a special type of adaptive design where items are selected in sets or modules. The simulation also contextualizes the principles of external item selection with a targeted test (TT) design. MST and TT designs are useful for illustration as these designs can appear to produce equivalent missing data structures, despite requiring distinct analysis methods. In both MST and TT designs, a proportion of items may be administered to all examinees, while subsets of items that differ in terms of difficulty are only administered to corresponding subsets of the sample that differ in proficiency. The simulation illustrates that the nature of how the examinees are routed to items, either based on information internal or external to the test, has important implications for unbiased item parameter estimation.

Data was generated with either an MST or a TT design, as described below. For both designs, the data generation was replicated 100 times. For each generated data set, two modeling approaches were taken. First, an IRT model was fit to the data where no grouping variables were specified. Second, an IRT model was fit to the data where routing group membership, g, was used as a grouping variable. Both models were fit with marginal maximum likelihood using the expectation-maximization algorithm (Bock & Aitkin, 1981).

The quality of the item parameter estimation was evaluated in terms of bias and Root Mean Square Error (RMSE). Bias was calculated as,

$$\text{bias} = \frac{1}{100}\sum_{i=k}^{100}\hat{t}_k - t \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{100}\sum_{i=k}^{100}\left(\hat{t}_k - t\right)^2} \tag{6}$$

where $\hat{t}_k$ is the estimated item parameter in the kth replication of the simulation, and t is the corresponding data generating value of the item parameter.

Multistage Testing Design

Data was generated based on a large-scale MST trial study. The study was a two-stage MST, and at both stages examinees (simulees) received a module of items. The first module was always a routing module. Based on performance on the routing module, examinees received an easy, medium, or hard module. There were multiple modules of each type. Within each module type (routing, easy, medium, or hard), the module was selected completely at random and at equal probability. In total there were three routing modules, two easy modules, four medium modules, and one hard module. Each module had 16 to 18 items, including both multiple-choice and constructed-responses items.

While the simulation was based on a large-scale MST trial study, the items in the MST trial study had also been used for a non-adaptive assessment. Item parameters estimated from the non-adaptive and non-targeted assessment were used as the population item parameters when generating the data. The three-parameter logistic model (3PLM) was used for multiple-choice items, the 2PLM was used for dichotomously-scored constructed-response items, and the generalized-partial-credit model (GPCM) was used for constructed-response items with three or more score categories.

A total of 60,000 test takers were simulated with $\theta \sim N(0, 1)$. Examinees were routed to the targeted modules based on *expected a priori* (EAP) scores estimated from the responses to multiple-choice routing items. Constructed-response items in the routing module were not used in the routing to simulate a realistic design, as such items were humanly scored after test completion in the original MST trial study.

The routing was based on EAP scores relative to two predefined thresholds, both of which were used in the original MST trial study. As a consequence, about 7,000 examinees were routed to easy modules, about 29,000 examinees were routed to medium modules, and about 24,000 examinees were routed to hard modules. Note that the exact method (EAP, thresholds) to calculate the routing decision is not relevant to the mathematical results in the present paper. The important feature is simply that the routing decision is based only on performance on the routing items.

Targeted Testing Design

Data was generated with the same item pool and item modules as in the MST simulation above, but with the TT design. Examinees received one of three targeted assessments, including either 1) a routing module and an easy module, 2) a routing module and a medium module, or 3) a routing module and a hard module. Examinees were selected into one of the three targeted assessments based on having low ($G = low$), medium ($G = medium$), or high proficiency ($G = high$), characterized by external information.

As with the MST simulation, there were multiple modules of each type that were selected at random with equal probability. Specifically, the same three routing modules, two easy modules, four medium modules, and one hard module, from the MST simulation were used. Note that routing module terminology is used only because the modules were also used in the MST design. The routing modules were not used for routing in the TT design.

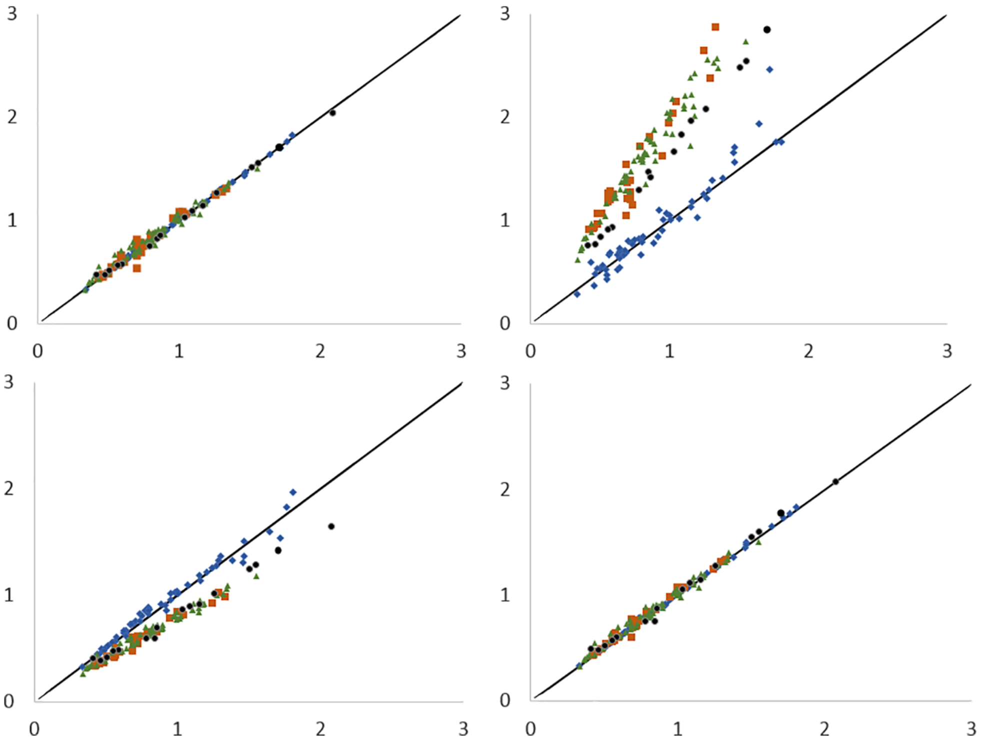
The population item parameters were obtained from estimates from a previous, non-adaptive and non-targeted assessment (see MST simulation section, above). To represent the use of external information in the item selection, θ was independently generated from three non-equivalent distributions depending on g. Specifically, $\theta | G = low \sim N(-1, .6)$, $\theta | G = medium \sim N(0, .6)$, and $\theta | G = high \sim N(1, .6)$. Data for 20,000 simulees were generated for each of three targeted groups. Overall, these specifications imply the total population has a mean of zero and a standard deviation of 1.

Results

Figures 1 and 2 show the mean estimated item parameters versus the generating values for the discrimination and location parameter of every item, respectively. The top left and bottom right panels of each figure show the results when the model is appropriate for the data: one group IRT for MST data and multigroup IRT for TT data, respectively. In contrast, the top right and bottom left panels of each figure show the results when the model is inappropriate for the data: one group IRT for TT data and multigroup IRT for MST data, respectively. As expected, when the model is appropriate for the data, no item parameter estimation bias was observed, but when the model is inappropriate, clear patterns of item-parameter estimation bias was observed.

Figure 1

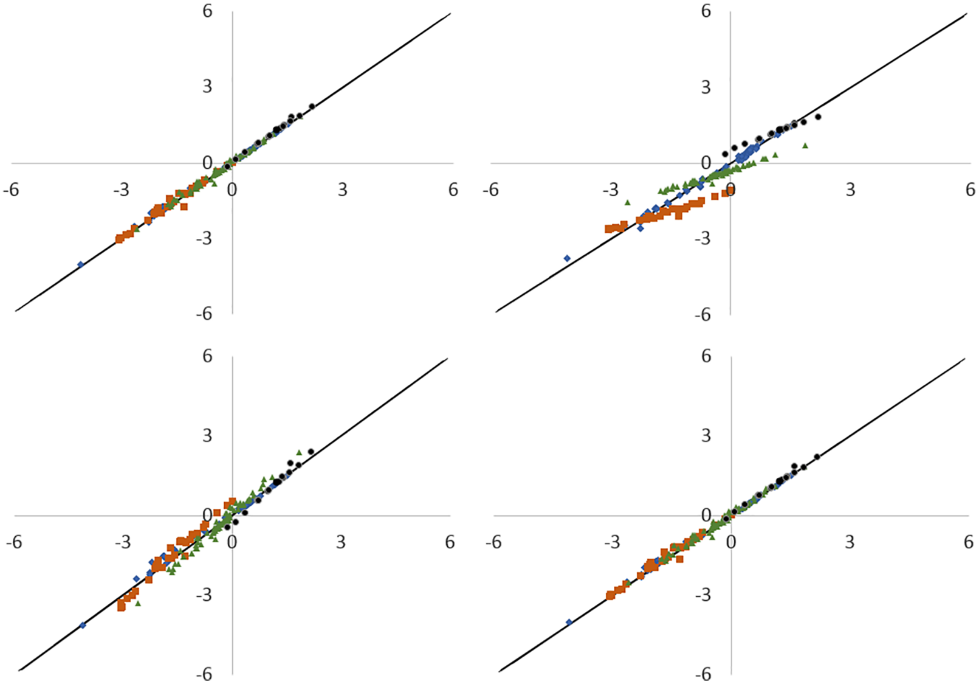
Discrimination Parameter Estimates Versus Generating Values



Note. Numbers on the Y-axis are the estimates, numbers on the X-axis are the generating values. Left panels are estimates with one group IRT, right panels are estimated with multigroup IRT. Top panels are based on MST data, bottom panels are based on TT data. Blue diamond = routing items, orange square = easy items, green triange = medium items, black diamond = hard items.

Figure 2

Difficulty Parameter Estimates Versus Generating Values



Note. Numbers on the Y-axis are the estimates, numbers on the X-axis are the generating values. Left panels are estimates with one group IRT, right panels are estimated with multigroup IRT. Top panels are based on MST data, bottom panels are based on TT data. Blue diamond = routing items, orange square = easy items, green triange = medium items, black diamond = hard items.

The mean bias and RMSE for the discrimination and location parameters are provided in Table 2. Because items from different module types (routing, easy, medium, and hard) show different patterns of bias, the bias and RMSE were averaged across all items within a given module type. Consistent with the figures, Table 2 shows no evidence of bias and smaller RMSEs when the appropriate model was used.

Table 2

Simulation Results for Multistage Testing and Targeted Testing Data

	Multistage Testing				**Targeted Testing**			
	One group IRT		**Multigroup IRT**		**One group IRT**		**Multigroup IRT**	
	bias	**RMSE**	**bias**	**RMSE**	**bias**	**RMSE**	**bias**	**RMSE**
a_r	0.00	0.03	0.04	0.09	0.02	0.06	0.01	0.04
a_e	0.02	0.09	0.72	0.74	−0.15	0.16	0.02	0.05
a_m	0.02	0.07	0.81	0.82	−0.16	0.17	0.03	0.07
a_h	−0.01	0.04	0.67	0.67	−0.20	0.21	0.00	0.06
b_r	0.03	0.06	0.00	0.09	0.08	0.11	0.04	0.07
b_e	0.03	0.13	−0.33	0.45	0.10	0.28	0.05	0.12
b_m	0.04	0.09	−0.10	0.32	0.00	0.21	0.05	0.10
b_h	0.02	0.04	0.05	0.18	−0.10	0.17	0.00	0.04

Note. MST = Multistage Testing (an example of internal item selection), TT = Targeted Testing (an example of external item selection), RMSE = Root Mean Square Error. *a* = discrimination parameter, *b* = location parameter. r, e, m, and h subscripts = routing, easy, medium, and hard, respectively. The numbers in the table are the mean bias and mean RMSE across all item parameters of the specified type.

Discussion

With internal item selection, the information used to select items is already accounted for in the estimation through the item responses, and the missing data is ignorable (Eggen & Verhelst, 2011; Glas, 1988; Jewsbury & van Rijn, 2020; Mislevy & Wu, 1996). Therefore, routing group membership does not need to be used as a grouping variable in MML estimation. Furthermore, because g was directly selected based on the item responses, including routing group membership as a grouping variable will erroneously make the equality assumption, so routing group membership must not be used as a grouping variable with internal item selection to ensure that the item parameter estimates are valid with MML.

With external selection, the information used to select items is not already accounted for the estimation of the item responses, so the missing data is non-ignorable. Including routing group membership as a grouping variable in the estimation is one way to appropriately account for the missing data (Mislevy & Sheehan, 1989). Unlike in internal item selection, routing group membership is not directly assigned based on the item responses in external item selection, so including routing group membership as a grouping variable does not necessarily introduce erroneous equality assumptions. For these reasons, routing group membership should be used as a grouping variable with external item selection data to ensure that the item parameter estimates are valid with MML.

There were clear directional patterns of bias in the simulation study when the inappropriate method was used for MST (an example of an internal item selection design) and TT (an example of an external item selection design). Notably, the biases were in opposite directions in the MST and TT simulations. Effectively, in the MST design when routing group membership was used as a grouping variable, the relative performance of the groups on the routing item responses was used twice, exaggerating the differences between the groups. In the TT design when routing group membership was not used as a grouping variable, some information determining differences between the groups was ignored in the estimation, causing the group differences to be under-represented. As data on targeted item responses is only observed for one routing group in MST and TT designs, these over- or understated differences are reflected in the item parameter estimates as bias.

While the present results for MML item parameter estimation in datasets generated by either adaptive or targeted testing designs are consistent with previous results in the literature (e.g., Eggen & Verhelst, 2011; Glas, 1988; Mislevy & Wu, 1996), the internal versus external routing distinction introduced in the present paper provides an intuitive and more general framework for applying MML and Bayesian inference. Furthermore, the framework also covers mixtures of internal and external item selection designs. Indeed, many MST designs involve some degree of external item selection and may be modeled inappropriately following usual guidelines for MSTs. For example, the MST design for the Programme for International Assessment of Adult Competencies (PIAAC) uses external information to select items in two ways (Chen et al., 2014; Yamamoto, Khorramdel, & Shin, 2018), in addition to internal information. First, examinees who report no familiarity with computers receive a non-adaptive test, while others receive an adaptive test. Second, the item selection in the MST is a function of not only performance on prior parts of the test, but also on the education level and native speaker status of the examinee. The results in this paper suggest that to ensure valid item parameter estimation, external information (computer familiarity, education level, and native speaker status) should to be accounted for in MML estimation while internal information (adaptive routing paths) should be ignored.

The internal versus external item selection framework also clarifies where assessment design principles may or may not generalize across different types of designs. Following standard practice in equating (von Davier, 2011), it has been suggested that a sufficient number of items must be common across test forms in adaptive designs to facilitate item parameter estimation to equate the test forms (e.g., Yamamoto, Shin, & Khorramdel, 2018, Yamamoto, Khorramdel, & Shin, 2018). However, while the need for a sufficient number of common items has been demonstrated in many equating problems (von Davier, 2011), these equating problems are examples of external item selection. With internal item selection, any differences between the examinees receiving different items can be fully explained by differences in item responses that are available for the model estimation.

Consequently, the design principle on the need for a sufficient number of common items does not apply to internal item selection designs.

In summary, internal and external item selection are defined by distinctive conditional (in)dependences between routing group membership, item responses, and θ (Equation 1, Equation 2, Equation 3, and Equation 4). Classification of item selection as either internal or external provides a simple, intuitive and general framework to inform item response theory analysis of datasets with missing data due to item selection.

Funding: This study was funded by ETS.

Competing Interests: The authors have declared that no competing interests exist.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics: Vol. 26. Handbook of statistics* (pp. 469–514). Elsevier

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Chen, H., Yamamoto, K., & von Davier, M. (2014). Controlling MST exposure rates in international large-scale assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 391–409). Chapman and Hall/CRC.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica, 32*, 107–132.

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics, 13*, 45–52. https://doi.org/10.3102/10769986013001045

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5–30. Springer

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT Models for Item Parameter Estimation with Multidimensional Multistage Tests. *Journal of Educational and Behavioral Statistics, 45*(4), 383–402.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices* (3rd ed.). Springer-Verlag.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society Edinburgh, Section A, 62*, 28–30. https://doi.org/10.1017/S0080454100006385

Lu, R., Jia, Y., & Wu, M. (2017, April 7). *Population definition and identification, priors, and non-random samples* [Conference presentation]. National Council of Measurement in Education Annual Meeting, San Antonio, TX, USA.

Lu, R., Jia, Y., & Wu, M. (2018, April 12). *Using design information in item parameter estimation with multistage testing* [Conference presentation]. National Council of Measurement in Education Annual Meeting, New York City, NY, USA.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177–185. https://doi.org/10.1007/BF02289699

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*, S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661–679. https://doi.org/10.1007/BF02296402

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

Muthén, B. O. (1989). Factor structure in groups on observed scores. *British Journal of Mathematical and Statistical Psychology, 42*, 81–90. https://doi.org/10.1111/j.2044-8317.1989.tb01116.x

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

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking*. Springer.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410605931

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wu, M., & Lu, R. (2017). *Multi-stage testing simulation studies.* Paper presented at the National Council of Measurement in Education annual meetingSan Antonio, TX.

Wu, M., & Xi, N. (2017 April 7). *Multi-stage testing in the 2015 NAEP mathematics DBA field trial* [Conference presentation]. National Council of Measurement in Education Annual Meeting, San Antonio, TX, USA.

Yamamoto, K., Shin, H., & Khorramdel, L. (2018). *Multistage adaptive testing design in international large-scale assessments. Educational Measurement: Issues and Practice., 37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yamamoto, K., Khorramdel, L., & Shin, H. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling, 60*, 347–368.

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika, 80*, 65–84. https://doi.org/10.1007/s11336-013-9369-6

Appendix

Theorem 1. *Under internal item selection, item parameter estimation with marginal maximum likelihood without routing-group membership g as a grouping variable in the model is valid.*

Proof. Under internal item selection, the marginal likelihood that is generally used by default, which ignores the missing data, is proportional to the actual marginal likelihood.

The marginal likelihood used when ignoring the missing responses is,

$$p(\tilde{\mathbf{y}}_{\text{obs}}|\boldsymbol{\beta}) \equiv \mathscr{L}^{*}(\boldsymbol{\beta}|\tilde{\mathbf{y}}_{\text{obs}}) = \int p(\tilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\beta})d\mathbf{y}_{\text{mis}},$$

where $\tilde{\mathbf{y}}_{\text{obs}}$ is the sample realization of the observed item response vector variable $\mathbf{Y}_{\text{obs}}$, $\mathbf{y}_{\text{mis}}$ is a vector of the item responses that are missing, $\boldsymbol{\beta}$ are the item parameters, and $\mathscr{L}^{*}$ is the assumed marginal likelihood.

Following Rubin (1976), the actual marginal likelihood is,

$$p(\tilde{\mathbf{y}}_{\text{obs}}, \tilde{\mathbf{m}}|\boldsymbol{\beta}, \phi) \equiv \mathscr{L}(\boldsymbol{\beta}, \phi|\tilde{\mathbf{y}}_{\text{obs}}, \tilde{\mathbf{m}}) = \int p(\tilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\beta})p(\tilde{\mathbf{m}}|\tilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \phi)d\mathbf{y}_{\text{mis}},$$

where ϕ are the parameters that govern the missing-data mechanism, $\tilde{\mathbf{m}}$ is the sample realization of the vector variable of missing data indicators $\mathbf{M}$, and $\mathscr{L}$ is the actual marginal likelihood.[1]

In internal item selection, $p(\mathbf{m}) = p(g)$, so the characteristic conditional independence in internal item selection (Equation 1), implies that the Missing At Random (MAR) assumption is satisfied. With the MAR assumption, the actual marginal likelihood can be simplified,

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}, \phi|\tilde{\mathbf{y}}_{\text{obs}}, \tilde{\mathbf{m}}) &= \int p(\tilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\beta})p(\tilde{\mathbf{m}}|\tilde{\mathbf{y}}_{\text{obs}}, \phi)d\mathbf{y}_{\text{mis}}, \\ &= p(\tilde{\mathbf{m}}|\tilde{\mathbf{y}}_{\text{obs}}, \phi)\int p(\tilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\beta})d\mathbf{y}_{\text{mis}}.\end{aligned}$$

Finally, with the assumption that $\boldsymbol{\beta}$ and ϕ is distinct (D), which is satisfied with internal item selection,

$$\mathscr{L}^{*}(\boldsymbol{\beta}|\tilde{\mathbf{y}}_{\text{obs}}) \propto \mathscr{L}(\boldsymbol{\beta}, \phi|\tilde{\mathbf{y}}_{\text{obs}}, \tilde{\mathbf{m}}),$$

which shows that marginal-likelihood inferences for $\boldsymbol{\beta}$, which are based on ratios of the likelihood function, are valid with internal item selection when the missing data is ignored. This conclusion is consistent with prior proofs with special cases of internal item selection designs (i.e., adaptive and multistage testing designs; Eggen & Verhelst, 2011; Glas, 1988; Mislevy & Wu, 1996). □

[1] Although y_{mis} is discrete in the context of IRT, we used integral notation to be consistent with Rubin (1976) and Mislevy and Wu (1996).

Theorem 2. *Under internal item selection, item parameter estimation with routing-group membership g as a grouping variable in the model is invalid.*

Proof. The assumed likelihood based on internal item selection data, when g is used as a grouping variable, is **not** proportional to the actual marginal likelihood.

The assumed marginal likelihood in a multigroup item response theory (IRT) model (Bock & Moustaki, 2007), is,

$$p(\mathbf{y}_{\text{obs}}|\boldsymbol{\beta}) \equiv \mathscr{L}^{*}(\boldsymbol{\beta}|\mathbf{y}_{\text{obs}}) = \prod_{g} \mathscr{L}^{*}(\boldsymbol{\beta}, g|\mathbf{y}_{\text{obs}}),$$

where,

$$\mathscr{L}^{*}(\boldsymbol{\beta}, g|\mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{c}}|\theta, \boldsymbol{\beta})p(\mathbf{y}_{g}|\theta, \boldsymbol{\beta})f(\theta|g)d\theta,$$

where $\mathbf{y}_{\text{c}}$ are the item responses to items observed in common for all examinees, $\mathbf{y}_{g}$ are the item responses to items only observed by group g, and θ is latent proficiency.

The actual marginal likelihood in a multigroup IRT model is,

$$p(\mathbf{y}_{\text{obs}}|\boldsymbol{\beta}) \equiv \mathscr{L}(\boldsymbol{\beta}|\mathbf{y}_{\text{obs}}) = \prod_{g} \mathscr{L}(\boldsymbol{\beta}, g|\mathbf{y}_{\text{obs}}),$$

where,

$$\mathscr{L}(\boldsymbol{\beta}, g|\mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{c}}|\theta, \boldsymbol{\beta}, g)p(\mathbf{y}_{g}|\theta, \boldsymbol{\beta})f(\theta|g)d\theta\,.$$

In internal item selection, $\mathbf{y}_{\text{c}}$ are used in the routing decision, and the characteristic conditional dependence in internal item selection (Equation 2) implies that,

$$p(\mathbf{y}_{\text{c}}|\theta, \boldsymbol{\beta}, g) \neq p(\mathbf{y}_{\text{c}}|\theta, \boldsymbol{\beta}),$$

and,

$$\mathscr{L}^{*}(\boldsymbol{\beta}|\mathbf{y}_{\text{obs}}) \not\propto \mathscr{L}(\boldsymbol{\beta}|\mathbf{y}_{\text{obs}}),$$

which shows that marginal-likelihood inferences for $\boldsymbol{\beta}$, which are based on ratios of the likelihood function, are invalid in internal item selection designs when the g is used as a grouping variable. □

Theorem 3. *Under external item selection, item parameter estimation without routing-group membership g as a grouping variable in the model is invalid.*

Proof. Under external item selection, the marginal likelihood that is generally used by default that ignores the missing data is **not** proportional to the actual marginal likelihood.

The marginal likelihood used when ignoring the missing responses is,

$$p(\tilde{\mathbf{y}}_{\text{obs}}|\boldsymbol{\beta}) \equiv \mathscr{L}^{*}(\boldsymbol{\beta}|\tilde{\mathbf{y}}_{\text{obs}}) = \int p(\tilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\beta})d\mathbf{y}_{\text{mis}},$$

where $\widetilde{\mathbf{y}}_{\text{obs}}$ is the sample realization of $\mathbf{Y}_{\text{obs}}$.

Following Rubin (1976), the actual marginal likelihood is,

$$p(\widetilde{\mathbf{y}}_{\text{obs}}, \widetilde{\mathbf{m}} \,|\, \boldsymbol{\beta}, \phi) \equiv \mathscr{L}(\boldsymbol{\beta}, \phi \,|\, \widetilde{\mathbf{y}}_{\text{obs}}, \widetilde{\mathbf{m}}) = \int p(\widetilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}} \,|\, \boldsymbol{\beta}) p(\widetilde{\mathbf{m}} \,|\, \widetilde{\mathbf{y}}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \phi) d\mathbf{y}_{\text{mis}},$$

where $\widetilde{\mathbf{m}}$ is the sample realization of $\mathbf{M}$. Under external item selection, $p(\mathbf{m}) = p(g)$, so the characteristic conditional dependence (Equation 4) implies that the MAR assumption is not satisfied. Because the MAR assumption is not satisfied, the actual marginal likelihood does not simplify, and,

$$\mathscr{L}^{*}(\boldsymbol{\beta} \,|\, \widetilde{\mathbf{y}}_{\text{obs}}) \not\propto \mathscr{L}(\boldsymbol{\beta}, \phi \,|\, \widetilde{\mathbf{y}}_{\text{obs}}, \widetilde{\mathbf{m}}),$$

which shows that marginal-likelihood inferences for $\boldsymbol{\beta}$, which are based on ratios of the likelihood function, are invalid in external item selection designs when the missing data is ignored. This conclusion is consistent with prior proofs with special cases of external item selection designs (i.e., targeted testing designs, e.g., Eggen & Verhelst, 2011; Mislevy & Sheehan, 1989). □

Theorem 4. *Under external item selection, item parameter estimation with routing-group membership g as a grouping variable in the model is valid.*

Proof. The assumed likelihood based on the external item selection data, when g is used as a grouping variable, is proportional to the actual marginal likelihood.

The assumed marginal likelihood in a multigroup IRT model (Bock & Moustaki, 2007), is,

$$p(\mathbf{y}_{\text{obs}} \,|\, \boldsymbol{\beta}) \equiv \mathscr{L}^{*}(\boldsymbol{\beta} \,|\, \mathbf{y}_{\text{obs}}) = \prod_{g} \mathscr{L}^{*}(\boldsymbol{\beta}, g \,|\, \mathbf{y}_{\text{obs}}),$$

where,

$$\mathscr{L}^{*}(\boldsymbol{\beta}, g \,|\, \mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{c}} \,|\, \theta, \boldsymbol{\beta}) p(\mathbf{y}_{g} \,|\, \theta, \boldsymbol{\beta}) f(\theta \,|\, g) d\theta\,.$$

The actual marginal likelihood in a multigroup IRT model is,

$$p(\mathbf{y}_{\text{obs}} \,|\, \boldsymbol{\beta}) \equiv \mathscr{L}(\boldsymbol{\beta} \,|\, \mathbf{y}_{\text{obs}}) = \prod_{g} \mathscr{L}(\boldsymbol{\beta}, g \,|\, \mathbf{y}_{\text{obs}}),$$

where,

$$\mathscr{L}(\boldsymbol{\beta}, g \,|\, \mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{c}} \,|\, \theta, \boldsymbol{\beta}, g) p(\mathbf{y}_{g} \,|\, \theta, \boldsymbol{\beta}) f(\theta \,|\, g) d\theta\,.$$

In external item selection, the characteristic conditional independence in external item selection (Equation 4), implies that,

$$p(\mathbf{y}_{\text{c}} \,|\, \theta, \boldsymbol{\beta}, g) = p(\mathbf{y}_{\text{c}} \,|\, \theta, \boldsymbol{\beta}),$$

and,

$$\mathscr{L}^*(\boldsymbol{\beta}|\mathrm{y}_{\mathrm{obs}}) = \mathscr{L}(\boldsymbol{\beta}|\mathrm{y}_{\mathrm{obs}}),$$

which shows that marginal-likelihood inferences for $\boldsymbol{\beta}$, which are based on ratios of the likelihood function, are valid in external item selection designs when the g is used as a grouping variable. □

https://doi.org/10.5964/meth.14823