

# Influence of Intersectional Routing Modules between Dimensions on Measurement Precision in Multidimensional Multistage Testing

Yi-Ling Wu and Yao-Hsuan Huang 

National Taiwan Normal University

Chia-Wen Chen 

University of Cambridge

Po-Hsi Chen

National Taiwan Normal University

*Multistage testing (MST), a variant of computerized adaptive testing (CAT), differs from conventional CAT in that it is adapted at the module level rather than at the individual item level. Typically, all examinees begin the MST with a linear test form in the first stage, commonly known as the routing stage. In 2020, Han introduced an innovative concept known as Intersectional Routing (ISR), which allows module selection in the first stage of the MST based on the examinees' estimated scores. These scores were predicted using a variety of information, including background data and other correlated latent traits.*

*In this study, we extend Han's ISR framework to a multidimensional test comprising multiple unidimensional subtests. In a multidimensional test, the correlation coefficients between the latent traits can be estimated by fitting a multidimensional item response theory model. Our extension allows module selection in the first stage of each subtest to consider information from all the other subtests via the known correlation matrix. The results of simulation studies showed that our extension improved the measurement compared with typical MST designs in conditions with moderate intercorrelations across module designs. The practical insights were given in the empirical analysis.*

## Introduction

Multistage testing (MST) has seen increased application in large-scale assessments (LSA) in recent years, such as the Programme for International Student Assessment (PISA) and Programme for the International Assessment of Adult Competencies (PIAAC) (Yamamoto, Shin, & Khorramdel, 2018). The PISA is designed for group-level reporting, and the PIAAC focuses on evaluating adult skills and knowledge. Additionally, other high-stakes assessments, such as the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), and Certified Public Accountants Examination (CPA), also utilize the MST mode (Breithaupt & Hare, 2007; Luecht et al., 2006).

Strategies for building computerized tests can be divided into two classes: linear tests, in which a fixed test form with the same items in the same order is shared

among multiple examinees (note that a linear test can contain multiple test forms for practical issues like item position balance); and adaptive tests, in which an individualized test form is created on-the-fly by the item selection based on the examinee's submitted responses. In computerized adaptive testing (CAT), individual items are sequentially administered (Weiss & Kingsbury, 1984), while in MST, sets of items named "modules" are the selection units (Zenisky et al., 2009). However, when CAT is employed for LSA, it faces constraints and difficulties in ensuring item-content balance, item-exposure control, and maintaining a wide range of difficulty item pools (Reese et al., 1999). The MST was introduced as a feasible alternative to CAT, offering a balanced compromise between linear tests and CAT (Hendrickson, 2007). Because the modules used in MST can be designed and assembled before test administration and are presented to the examinees as a unit, they allow test developers greater control over the content balance, quality of the test structure, and administration of the test in comparison to CAT, while maintaining the measurement advantages of CAT (van der Linden & Glas, 2010). Hambleton and Xing (2006) demonstrated that an MST design can make better use of item banks for a wider range of difficulty, and that content specifications are easier to meet.

Utilizing an MST design in LSAs offers at least two advantages over CAT. First, since experts can assemble MST test forms prior to administration, it better meets the design constraints and specific goals of LSAs than CAT (Yamamoto et al., 2018). These goals include maintaining content balance across all proficiency levels and adhering to item position constraints. MST employs a balanced incomplete block (BIB) design within its item modules, enabling experts to review predesigned test forms and prespecify item positions to mitigate potential effects related to item orders. This control allows for improved coverage of construct frameworks. For instance, PISA 2018 implemented a BIB design for the linear-tested mathematics and science domains, which balanced the item position effect. The combinations of domains in PISA 2018 also used a sort of BIB design, which effectively balanced domain position effects (Yamamoto et al., 2019).

Second, MST facilitates the partial incorporation of open-ended response items that cannot be automatically scored in LSAs. It can adaptively select the next item module based solely on automatically scored responses when mixed with items that require manual scoring.

While adaptive testing is primarily aimed at enhancing measurement precision for individuals, simulation studies have demonstrated that improving measurement accuracy for each test taker can also enhance overall group measurement accuracy (Tang et al., 2024). Consequently, MST designs are not only appropriate but also straightforward to implement in the context of group-score assessments such as LSAs (Yamamoto et al., 2018).

Most LSAs are consistent with a simple-structure multidimensional item response theory (MIRT) model in which each item measures only one latent variable, including the use of separate unidimensional item response theory (IRT) models for each latent variable (Mislevy et al., 1992; von Davier et al., 2006). Han (2020) proposed a framework for developing an MST with Intersectional Routing (ISR), enabling module selection in the first stage of the MST based on the examinees' estimated scores. These scores were predicted from various types of information, including back-

ground data and other correlated latent traits, with the potential to enhance measurement efficiency and test optimality. Consequently, we extend Han's ISR framework to a multidimensional test composed of multiple unidimensional subtests. When a multidimensional test is employed with known correlation coefficients between latent traits, the MST framework with ISR can be applied based on previously estimated scores. This study aims to examine the impact of different ability estimation approaches and combinations on the measurement precision of multidimensional MST under different panels, utilizing MST in conjunction with the ISR framework.

We took PIAAC as an example to demonstrate how ISR can be applied in MSTs within the context of LSAs. The PIAAC Literacy and Numeracy assessments followed an adaptive MST design, with IRT item parameters calibrated in a prior field test. Consequently, this represents an MST with precalibrated item parameters (Yamamoto et al., 2018). Participants completed either the Literacy assessment before Numeracy or vice versa. The ISR design can be applied to the PIAAC MST by modifying the module selection process for Literacy and Numeracy assessments to incorporate their interrelationship. Specifically, ISR leverages the correlation between Literacy and Numeracy scores, obtained from the precalibrated item parameters within a multidimensional IRT framework. In this approach, the selection of Literacy modules in the ISR design considers not only background information but also Numeracy scores, and vice versa. This study presents an equation that facilitates the practical implementation of the ISR design and conducts simulation studies to evaluate its effectiveness in enhancing measurement precision.

## Literature Review

### MIRT

Measuring latent traits using item parameters and examinees' item responses has been a successful application of IRT (Lord et al., 1968). The two fundamental assumptions supported by IRT models are unidimensionality and local independence (Hambleton & Swaminathan, 1985; Lord, 1980). Unidimensionality assumes that each item measures a single latent trait. However, MIRT allows the simultaneous estimation of item parameters for multiple dimensions, enabling the measurement of multiple latent traits. Notably, the IRT and MIRT models are based on different assumptions regarding the latent dimensions.

To measure multiple latent traits, the MIRT model provides more precise latent trait estimates than the IRT model by utilizing the correlation coefficient matrix between latent traits. Figure 1 illustrates the structural differences between the IRT and MIRT models. IRT assumes a zero correlation between traits, whereas MIRT concurrently calibrates item parameters and the variance-covariance of multiple traits in model estimation. When applying CAT and MST, where the calibrated parameters are known, trait estimates in the MIRT framework consider the covariance between traits, which is not the case in the IRT framework.

### Significance and Contribution of MST Design

Recently, MSTs have become increasingly popular, and can be considered as special cases of CAT. The design of an MST is defined by the number of sequential

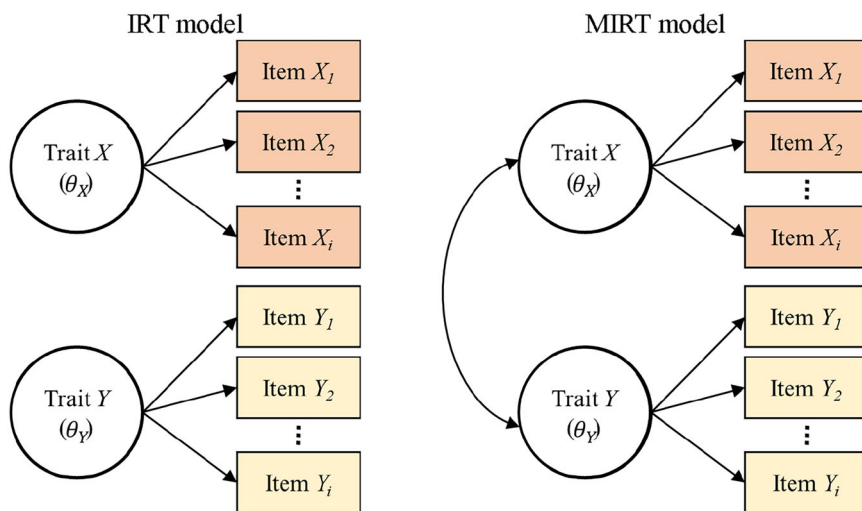


Figure 1. Structures of the IRT and MIRT models.

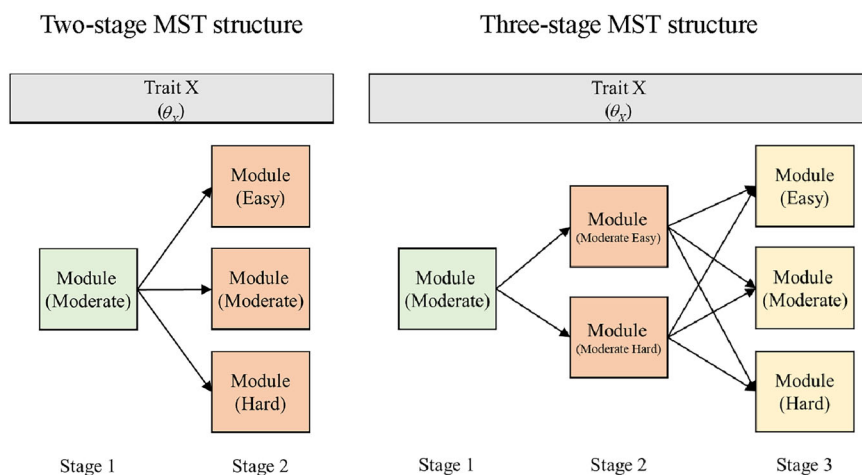


Figure 2. Two-stage MST (left) and three-stage MST (right). *Note:* In the two-stage example, the first stage consists of one module with moderate difficulty and the second stage consists of three different difficulty modules. In the three-stage example, the first stage consisted of one moderate module, the second stage consisted of two difficulty modules, and the third stage consisted of three modules with different difficulties.

stages, modules, and paths that can be followed between stages. A module is a collection of test items to be administered at each stage. It is common for several modules to be preassembled for each stage to form a panel. Figure 2 shows a side-by-side comparison of two- and three-stage MST design structures. In the initial stage, there is typically a routing module, whereas in the subsequent stages, there are separate modules at different levels, such as easy, moderate, and hard difficulty modules. The test forms are preconstructed into a structure (e.g., panel 1-2-3 on the right

of Figure 2) with modules and paths between stages; the administration algorithm routes each examinee from one module to the next based on the structure, and the examinee's ability is estimated at the end of each module (Luecht et al., 2006, Luecht & Nungester, 1998). MST procedures and algorithms, such as module selection and ability estimation, have been developed and implemented based primarily on IRT. MST provides more flexibility to test developers, allowing mixed item formats, such as constructed-response items and testlets (Zenisky et al., 2009). It overcomes some of the problems associated with CAT, with a slight effect on test efficiency (Patsula & Hambleton, 1999; Zenisky et al., 2009). For example, response review is an issue. Although CAT systems typically limit examinees from reviewing their responses (Hambleton et al., 1991; Vispoel, 1998; Vispoel et al., 2000; Wainer, 2000; Yen, 1993), it is technically feasible to conduct a response review in CAT. However, implementing such a provision may affect the optimality and efficiency of the item selection algorithm. While some studies permit response reviews in CAT (Han, 2013; Stocking, 1997; Wang et al., 2017), significant drawbacks exist in the biased scores of participants who strategize by first answering everything wrong (so they get the easiest test) and then revising all their initial answers (Wainer, 1992). Also, the missing-at-random assumption in CAT is broken when response review is allowed, which results in a biased estimation of item parameter updates in the item bank using CAT response data under response review (van der Linden & Glas, 2010). Notably, in contrast to CAT, MST permits examinees to review their item responses within each module with missing responses under the missing-at-random assumption.

MST has been successfully applied to high-stakes assessments, such as GRE, TOEFL, and CPA (Breithaupt & Hare, 2007; Luecht et al., 2006), as well as the international LSAs, such as PIAAC and PISA (Yamamoto et al., 2018). Educational assessments often have multidimensional structures. Even in cases with relatively simple latent structures, each item typically measures only one latent trait in a multidimensional test. For example, PISA 2018 included three tests with a simple latent structure. Each test measured one of three latent traits: reading, scientific, and mathematical literacy. The subtests are administered according to an integrated design in which students take either two or three subtests (see PISA 2018 Tech Report, figure 2.5). Conceptually, when tests are administered sequentially, latent trait estimation in the later administered test can benefit from information from the earlier administered test. This is based on the prior information of the factor covariance matrix as the MIRT is applied. However, currently, testing bodies adopt multiple unidimensional MST designs for a test with a simple multidimensional structure. Specifically, MST designs for measuring different latent traits are applied independently of each other, ignoring correlations between traits, even within the same assessment program (e.g., PISA and PIAAC). In the CAT context, the reliability of ability estimation in a multidimensional structure can be improved by introducing the correlation between domains under the MIRT framework compared to a multiple-unidimensional IRT framework where multiple traits were independent of each other in ability estimation and item selection (Wang & Chen, 2004). More statistical details about utilizing intertraits correlation to improve each of multiple unidimensional ability estimation precisions are written in Section 2.4.

Wu et al.

Multidimensional adaptive testing (MAT) is a highly efficient method for the simultaneous measurement of multiple latent traits. Han (2020) showed that the use of MAT in the MST approach could improve measurement efficiency and test optimality, particularly for short tests. In this study, the MIRT model was used in the context of MST. The model is expressed as

$$P(X_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{e^{\mathbf{a}_i^T \boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i^T \boldsymbol{\theta}_j + d_i}} \quad (1)$$

where  $P(X_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j)$  is the probability of a correct response for examinee  $j$  to item  $i$ ,  $X_{ij}$  is the item response for examinee  $j$  to item  $i$ ,  $\mathbf{a}_i$  is a vector of parameters related to the discrimination of item  $i$ ,  $d_i$  is a parameter related to the intercept of item  $i$ , and  $\boldsymbol{\theta}_j$  is a vector of latent traits for examinee  $j$  (McKinley & Reckase, 1982).  $\boldsymbol{\theta}_j$  follows a multivariate normal distribution with a mean vector  $\boldsymbol{\mu}$  and variance-covariance matrix  $\boldsymbol{\Sigma}$ . The model is typically identified by fixing the means at zero and the variances at one, so that the off-diagonal of  $\boldsymbol{\Sigma}$  contains the correlations between the latent traits.

### Routing Strategy in MST Design

The MST strategy assigns examinees to the next module based on their performance in the previous module, similar to the CAT item-selection procedure. An IRT framework with information-based objectives can be used to implement criterion-based routing (Luecht & Nungester, 1998; Weissman, 2014; Zenisky, 2004). The approximate maximum information method (Luecht et al., 2006) has been widely adopted. The procedure involves summing the test information from a previously delivered module and current alternative modules and finding significant test information. Based on their performance in the previous stages, the examinees were routed to one of the modules in each stage. Considering the current preliminary estimate, this procedure is equivalent to the maximum information item selection strategy for CAT.

The module items in the first stage were fixed in typical MST designs. However, this constraint can be lifted, allowing for an adaptive selection procedure for examinees if they have access to information related to the current measured trait before test administration. Han (2020) proposed an ISR framework to adaptively assign a module in the first stage of MST. This approach significantly enhances the assessment adaptability and improves the measurement efficiency. This is similar to the initial item selection process in CAT, in which the initial item choice aims to minimize the standard error of the initial latent trait estimates. This was achieved by utilizing various sources of information, including prior test administration results, scores from other programs measuring similar latent traits, data from learning systems (e.g., school population means), and response time data (Thompson & Weiss, 2011; van der Linden & Pashley, 2010). An example of utilizing the background information to improve latent trait estimation can be found in PIAAC, where the plausible values of the latent scores were given after controlling for the effect of students' demographic variables (Khorramdel et al., 2020).

Generally, a moderate-to-high correlation exists between test section scores representing different traits. For instance, many educational assessments show a

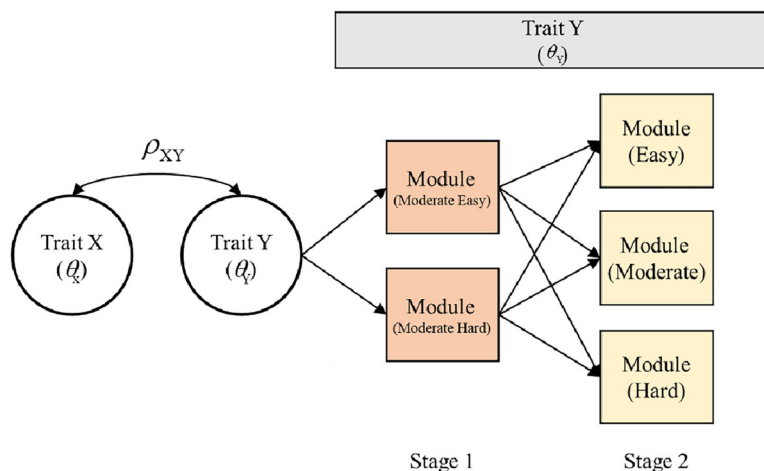


Figure 3. MST design with ISR (Han, 2020).

correlation coefficient ranging from .40 to .70 between math and language section scores (Ding & Homer, 2020; Peng et al., 2020; Ünal et al., 2023). Hence, when we have information on an examinee's score for a correlated trait and knowledge of the relationship between these traits, we can predict the score for the currently measured trait in advance using the MIRT model. In MST, this adaptive routing based on the predicted section score is referred to as ISR, as shown in Figure 3.

The initial estimates  $\theta_Y$  for selecting the module in the first stage of MST by measuring dimension  $Y$  were obtained by utilizing information on a known latent trait estimate of trait  $X$  ( $\theta_X$ ) and the correlation between traits  $X$  and  $Y$  ( $\rho_{XY}$ ). Han (2020) proposed an ISR framework that could increase MST measurement efficiency and reduce the test length by 10% without causing score estimation bias. In this study, we simplified and extended the ISR framework to a multidimensional MST design using a simple multidimensional structure.

### Estimate Latent Trait

Using Bayesian-based estimation methods in a multidimensional test can incorporate the intertrait covariance into ability estimation to improve the measurement precision and reliability (Wang & Chen, 2004). Maximum Likelihood Estimation (MLE) is a commonly used method to estimate latent traits in IRT. However, MLE is an outward-biased estimator for ability estimates (Warm, 1989) and cannot handle information, including those that consist only of all correct or incorrect responses, especially in the initial stages of administering MST, increasing the chance of extreme response patterns. We used Bayesian-based estimation methods (Owen, 1975) to handle the extreme response patterns, such as the modal a posteriori (MAP) method (Samejima, 1969) or the expected a posteriori (EAP) method (Bock & Aitkin, 1981). The MAP and EAP methods have a peak from which a score estimate can be obtained by imposing a prior distribution on the log-likelihood function (Segall, 1996).

However, the MAP method employing the iterative Newton-Raphson process has nonconvergence problems. The EAP method does not employ the iterative Newton-Raphson process but rather computes the conditional posterior distribution for each quadrature point to obtain the mean of the posterior distribution. Empirically, it requires a significant computation time when the number of dimensions in a test is large. Despite the conceptual and mathematical differences between EAP and MAP, unless the mode and mean of a posterior distribution differ significantly, EAP and MAP often produce similar estimation results (Han, 2016). This study employs the EAP method for ability estimation because we assume that the prior distribution used in this study approximates the real population fairly. Provided that the number of quadrature points is set effectively, utilizing the EAP method for ability estimation can achieve accurate results.

In this section, we describe the EAP for multidimensional ability estimation (Segall, 1996; Reckase, 2009). Suppose that examinee  $j$  responds to a set of items with item score vector  $X_j$ , the likelihood of observed responses  $X_j$  given multidimensional ability  $\theta$  is written as  $L(X_j|\theta) = L(X_{1j}, X_{2j}, \dots|\theta) = \prod_{i \in S} P(X_{ij}|\theta)^{x_{ij}} Q(X_{ij}|\theta)^{1-x_{ij}}$  where  $P(X_{ij}|\theta)$  is defined in Equation 1,  $Q(X_{ij}|\theta) = 1 - P(X_{ij}|\theta)$ , and  $S$  is a vector containing the administered items. Under the Bayesian theorem, the posterior distribution of  $\theta$  is expressed as  $f(\theta|X_j) = L(X_j|\theta) \times \frac{f(\theta)}{f(X_j)}$  where  $f(\theta)$  is the prior distribution of  $\theta$ , which was assumed to follow a multivariate normal distribution with means of  $\mu$  and covariance matrix of  $\Phi$ , and  $f(X_j)$  is a marginal probability of  $X_j$  given by  $\int_{\theta} L(X_j|\theta) \times f(\theta) d\theta$ . The MAP estimator is finding the solution of maximizing the log posterior function, that is  $\ln f(\theta|X_j) = L(X_j|\theta) - \frac{1}{2}(\theta - \mu)' \Phi^{-1}(\theta - \mu) + C$ , where  $\ln$  is the natural logarithm function, and  $C$  is a constant that can be ignored in determining the maximum of the function (Segall, 1996). The log posterior function contains the intertrait covariances in  $\Phi$ , which statistically increases the empirical test reliability of each dimension (Wang & Chen, 2004). The EAP estimate is the mean of the posterior distribution expressed as  $E(\theta|X_j) = \frac{\int_{\theta} \theta \times L(X_j|\theta) \times f(\theta) d\theta}{\int_{\theta} L(X_j|\theta) \times f(\theta) d\theta}$ . The integration calculation can be approximated using a weighted sum over a set of integration points (i.e., quadrature nodes) or the Monte-Carlo integration approach by adding randomly generated integration points.

Under a simple multidimensional structure, suppose  $c$  is the vector to denote the dimensions that an examinee has completed, and  $\theta_c$  contains elements of ability estimates for such dimensions. Let  $\theta_h$  denote the ability of the current dimension that the examinee is going to assume but has not yet taken. According to the general theory of multivariate normal distributions (Eaton, 1983, pp. 116–117), the  $\theta_h$  estimate is the expected value of a conditional normal distribution  $f(\theta_h|\theta_c)$  with a mean equal to

$$\mu_{h|c} = \mu_h + \Sigma_{hc} \Sigma_{cc}^{-1} (\theta_c - \mu_c) \quad (2)$$

and variance equal to  $\sigma_{h|c}^2 = \sigma_h^2 - \Sigma_{hc} \Sigma_{cc}^{-1} \Sigma'_{hc}$ , where  $\mu_c$  and  $\Sigma_{cc}$  are the mean vector and covariance matrix of abilities  $\theta_c$ ,  $\Sigma_{hc}$  is the vector of covariance between  $\theta_h$  and  $\theta_c$ , and  $\mu_h$  and  $\sigma_h^2$  are the marginal mean and variance of  $\theta_h$ . This study used

Equation 2 to obtain the estimate of  $\theta_h$  equal to  $\mu_{h|c}$ , for the dimension  $h$  that an examinee is going to assume after completing dimension  $c$ .

### Application of ISR Framework in MST in LSAs Context

The applications of MST in LSAs can either be done with precalibrated item parameters (Min & Bishop, 2024) or with the post hoc calibration in LSAs that estimates item parameters from the data collected through the MST designs (Yamamoto et al., 2018; Yamamoto et al., 2019). Traditionally, MST item parameters are precalibrated, where all item parameters are known through pilot studies. In both cases (i.e., precalibrated and post hoc calibration), implementing ISR requires prior information about the correlation matrix between traits and the trait scores other than the current test domain. For example, although PISA 2018 uses a post hoc calibration and only one domain (reading test) is involved in MST design, in such cases, the ISR can still improve the ability estimation by utilizing (1) the scores in other domains administered before the currently taken domain, regardless of whether the previous domains are in an MST design or (2) the noncognitive scores correlated to the current domain, and (3) the correlation matrix between all traits based on PISA 2015 or even other LSAs results. Specifically, the reading domain MST can adopt ISR design by incorporating science and math scores if one or both have been administered to the test taker. Also, the self-efficacy scores can be incorporated together with science and math test scores to implement ISR by using correlations between those traits we have assumed based on PISA 2015. The ISR allows us to select a module at the beginning of MST by Equation 2 as long as we have domain scores correlated to the current domain and the prior correlations between domains. When we obtain more trait scores or higher correlated trait scores for a test taker before she/he is currently taking a domain test, ISR gives a better choice of test module because more information is utilized for current ability estimation.

### Research Purpose

This study aims to simplify and extend the ISR framework to a multidimensional MST design with a simple latent structure composed of multiple unidimensional subtests. We investigated the effects of the test length, number of candidate modules in the first stage, and item bank size on the precision of trait estimation for simplified ISR compared with typical multiple unidimensional MSTs by conducting simulation studies. We hypothesized that the simplified ISR approach would perform better than typical MSTs in terms of the precision of trait estimation when the correlation between traits is moderate to high.

### Multidimensional MST with ISR Structure

Similar to MAT, which extends CAT to estimate abilities and maximize test information in a multidimensional framework, ISR is an extension of MST from a unidimensional to a multidimensional framework in terms of ability estimation and module selection (Han, 2020). Unless there is zero correlation among the measured latent traits, a known covariance matrix can be utilized as a prior for estimating the score on the currently measured dimension based on another dimension in MAT

using Bayesian-based estimation methods, even when no item measuring the dimension being estimated is administered in the initial stage of the test. Specifically, after the examinees complete one or more subtests, the module selection in the first stage of the subsequent subtest is based on the initial ability estimates, which are obtained using information from all other administered subtests and the known correlation matrix. Figure 4 shows a multidimensional MST with an ISR structure and a typical multidimensional MST. The upper panel in Figure 4 shows a typical MST, where the first stage of all the subtests consists of fixed tests. The multidimensional MST with ISR in the lower panel of Figure 4 shows that after the first subtest was administered, the first stage of the subsequent subtests became an adaptive selection procedure.

For obtaining the initial trait estimates of the subsequent tests  $\hat{\theta}_Y$  when the first subtest for trait  $X$  has been administered, Han (2020) suggests the following steps:

1. Conduct a Monte-Carlo simulation for the two subtests to generate an item response matrix, where the individuals' true values of  $\hat{\theta}_X$  and  $\hat{\theta}_Y$  followed a multivariate normal distribution with the correlation known from the literature or any practical resources.
2. Estimate the individuals' trait scores  $\hat{\theta}_X'$  and  $\hat{\theta}_Y'$  for the two subtests by analyzing the data generated in Step 1 in the MIRT model.
3. Conduct a regression to predict the latent scores for trait  $Y$  by the latent scores for trait  $X$ , that is  $\hat{\theta}_Y' = \beta_0 + \beta_1 \hat{\theta}_X' + \varepsilon$ , and estimate the regression coefficients  $\beta_0$  and  $\beta_1$ .
4. Estimate the latent score  $\hat{\theta}_X$  for Trait  $X$  based on the observed responses to the first subtest in the IRT model.
5. Use  $\hat{\theta}_X$ ,  $\beta_0$ , and  $\beta_1$  to predict the initial estimate  $\hat{\theta}_Y$  by regression.

This study proposes an ISR with a simplified routing procedure based on prior subtest results and prior trait distributions using MIRT modeling in a simple latent structure. The following steps are suggested:

1. Concurrently estimate  $\hat{\theta}_X$  and  $\hat{\theta}_Y$  in a Bayesian estimator (e.g., MAP or EAP and Equation 2) with the prior trait distribution and a known correlation matrix when the first subtest has been administered.

An apparent difference between Han's ISR and the simplified ISR is that the number of steps was reduced from five to one. Second, the simplified ISR avoids the Monte-Carlo simulation in Han's first step, which may be difficult to implement for practitioners who do not have experience in data generation. Although the simplified ISR uses a Bayesian estimator that is more difficult for practitioners than regression, practitioners can still implement the simplified ISR without understanding the Bayesian estimator because the existing software (e.g., package *mirt* in R) can run it by feeding the observed response data, model specification, and prior distribution into the software function.

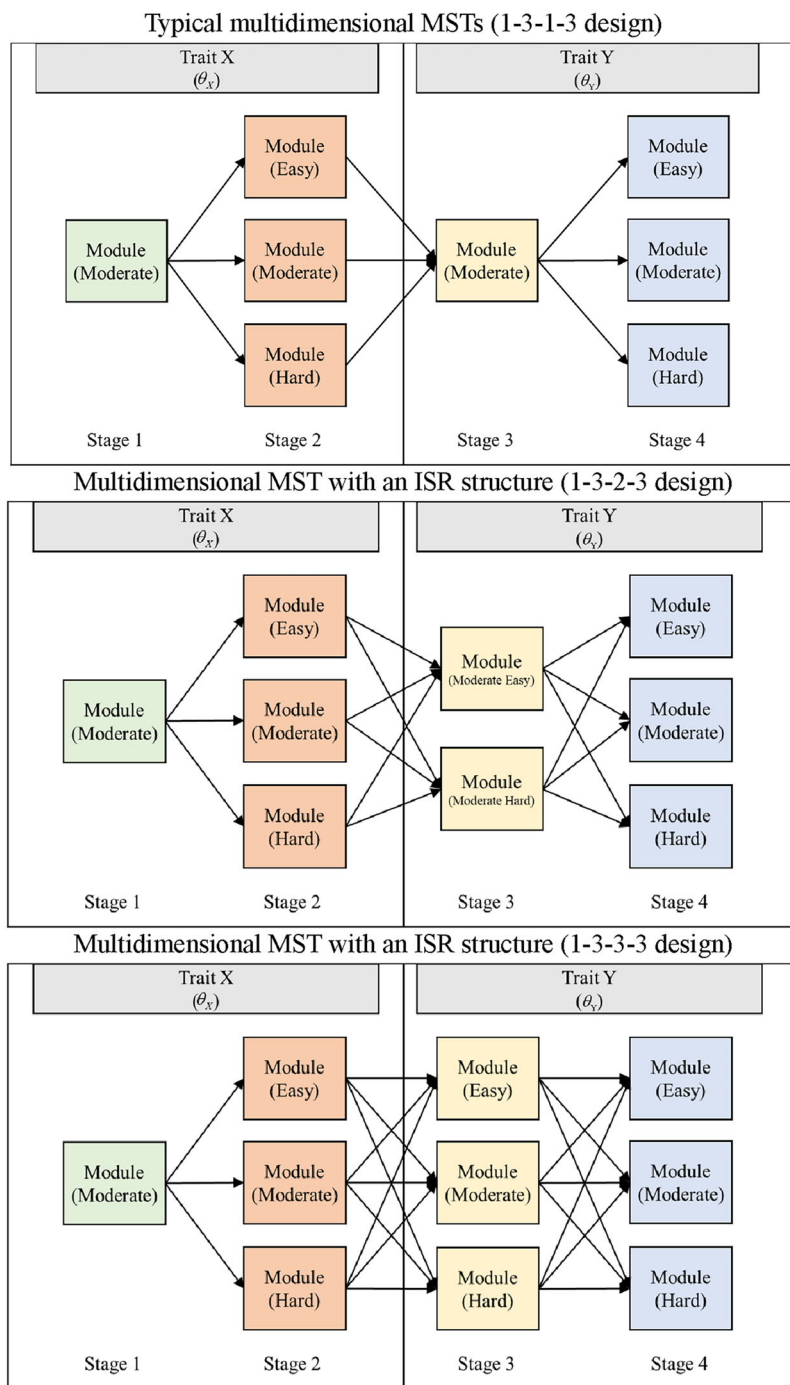


Figure 4. Multidimensional MST with an ISR structure.

Table 1  
*Summary of Item Parameters*

Module	Discrimination	Easiness	Interval of Easiness
Easy	$a_{i1} \sim N(1, .2)$	$d_i \sim N(1, 1)$	$(.5, \infty)$
Moderate easy	$a_{i2} \sim N(1, .2)$	$d_i \sim N(.5, 1)$	$(0, \infty]$
Moderate		$d_i \sim N(0, 1)$	$(-.5, .5]$
Moderate hard		$d_i \sim N(-.5, 1)$	$(-\infty, 0]$
Hard		$d_i \sim N(-1, 1)$	$(-\infty, -.5]$

*Note.* Curves and square brackets in the ease interval indicate open and closed intervals, respectively.

Table 2  
*Summary of the Number of Items in Each Panel and Item Banks*

	Test Length of the Whole Test		
	40	60	80
Size of each module	10	15	20
Test length of each subtest	20	30	40
Item bank of panel 1-3-1-3	80	120	160
Item bank of panel 1-3-2-3	90	135	180
Item bank of panel 1-3-3-3	100	150	200

### Simulation Study

This section discusses the simulation study conducted to examine the performance and effectiveness of implementing ISR in MST for estimating latent traits. This study explored scenarios involving one, two, or three modules for routing, referred to as panels 1-3-1-3, 1-3-2-3, and 1-3-3-3, respectively. Panel 1-3-1-3 shows the typical multidimensional MST design, and panels 1-3-2-3 and 1-3-3-3 show the multidimensional MSTs with ISR, as shown in Figure 4. Two-dimensional latent traits were sampled from a multivariate normal distribution with correlations of .3, .5, or .7, consistent with Han's study (2020).

Table 1 shows a summary of item parameter generation. Easy, moderate, and hard modules were used in the MST design, as shown in Figure 4. When a stage contained three modules for selection, the easy, moderate, and hard models were included in the item bank. When a stage contained two modules for selection, moderately easy and moderately hard modules were included in the item bank. A moderate module was used when only one module was on the stage. Table 2 summarizes the number of items in each panel and the item bank. We manipulated the test length under the three conditions by manipulating the number of items for each module set to 10, 15, or 20. For example, when the number of items in a module is set to 10, a total of eight modules are contained in a test with panel 1-3-1-3, and the whole size of the item bank is 80. The examinees were given only two modules of 20 items in each subtest, resulting in a total of 40 items for each examinee.

In summary, we manipulated three simulation factors: the number of modules (1-3-1-3, 1-3-2-3, and 1-3-3-3), the correlation between two traits (.3, .5, and .7), and the test length (10, 15, and 20 items in each stage). We, therefore, had 3 (number of modules)  $\times$  3 (intertrait correlations)  $\times$  3 (test lengths) = 9 simulation conditions. We manipulated the number of modules because we expected more modules in the item bank to be better for measurement precision in MST design, which corresponds to the findings of the literature (Patsula, 1999). We manipulated the intertrait correlation because we expected the larger intertrait correlations, the better performance of ISR. High-correlated traits that ISR utilized in Equation 2 can largely explain the variation of the currently administered trait and reduce the variance of the posterior distribution in EAP estimation. We manipulated the test length because Han (2020) declared that the ISR design can improve measurement precision when the test lengths are short. The advantage of ISR became limited in the context of long test lengths since ISR utilizes only prior information in module selection, of which the effect is diluted when the number of items increases.

The sample size was 10,000 each with 100 replications. All processes related to data generation, research procedures, and results analysis were conducted using the software packages *mirt* (Chalmers, 2012) and *mirtCAT* (Chalmers, 2016) within R software (v4.1.3; R Core Team, 2021). The pseudo codes in R can be found in the Appendix for the simulation of MST with ISR.

### Evaluating the Panels

The bias and root mean square error (RMSE) of the estimated latent traits with respect to comparison values for individual  $j$  were evaluated in the simulation study. Specifically,

$$bia s_j = \frac{1}{R} \sum_{s=1}^R (\hat{\theta}_{js} - \theta_j), \quad (3)$$

$$RMS E_j = \sqrt{\frac{1}{R} \sum_{s=1}^R (\hat{\theta}_{js} - \theta_j)^2} \quad (4)$$

where  $\hat{\theta}_{js}$  is the latent trait estimate for examinee  $j$  and replication  $s$ ,  $\theta_j$  is the comparison value of the latent trait parameter for examinee  $j$ , and  $R = 100$  is the number of replicates. In this simulation study, comparison values were used to generate the true values. We reported the mean and standard deviation of bias and RMSE across all 10,000 people.

Furthermore, we evaluated the reliability of the replication  $s$  given by the squared correlation between the latent trait estimates and the generated true values. The mean and standard deviation of the reliability across 100 replicates are reported for each simulation condition. In addition, we evaluated the conditional bias for fixed specific ability levels equal to  $-2.5, -2.0, -1.5, -1.0, -.5, .0, .5, 1.0, 1.5, 2.0$ , and  $2.5$  for the second-dimension  $\theta_\gamma$ .

## Results

Tables 3–5 present the bias, RMSE, and reliability of latent trait estimation across the studied conditions. As shown in Table 3, bias remained consistent across all study conditions. However, the RMSE (as seen in Table 4) exhibited a tendency to decrease as the correlation  $\rho_{XY}$  increased in the ISR conditions. This aligns with our expectation that a higher correlation between latent traits allows more information about subsequent latent traits to be derived from the administered latent traits.

Compared with the number of routing modules in Stage 3 (i.e., comparison among panels 1-3-1-3, 1-3-2-3, and 1-3-3-3), the RMSE decreased as the number of routing modules increased. This result is reasonable, because the large bank size in Stage 3 should improve the precision of the latent trait estimation. Compared to the test length of each dimension, the RMSE decreased as the test length of each dimension increased. A longer test length led to a lower RMSE and higher precision in latent trait estimation.

The reliability coefficient increased among the ISR conditions as intertrait correlation increased (Table 5). This trend in reliability yielded the same conclusions as those for the RMSE. In particular, the high  $\rho_{XY}$  can enhance the efficiency of the ISR approach owing to increased shared information between dimensions. Similarly, when comparing the number of routing modules in Stage 3, the reliability increased as the number of routing modules increased. The reliability tended to increase as the test length of each dimension increased. These results align with our expectation that a large bank size and test length can improve the reliability of latent trait estimates.

Conditional bias ranged from  $-.8$  to  $.8$  for all levels of latent traits. It showed a higher bias for examinees with extreme abilities and nearly no bias for those with intermediate abilities, as shown in Figures 5 and 6 for  $\theta_Y$ . The longer the test length, the more unbiased the estimation across panels. Having two modules in stage 3 improved the effectiveness of MST with ISR at extreme abilities (e.g., ability levels of  $-2$  and  $2$ ). However, having three modules in Stage 3 performed the best at moderate levels of ability. Specifically, panel 1-3-3-3 proved to be more effective than panel 1-3-2-3 and even more than panel 1-3-1-3. Positive biases for low-ability levels and negative biases for high-ability levels indicate a bias toward the center of a prior distribution, a well-known characteristic of Bayesian estimators (Lord, 1986).

We observed slight differences in the RMSEs and reliability among panels 1-3-1-3, 1-3-2-3, and 1-3-3-3 at the end of the MSTs. Thereafter, we evaluated the change in RMSE and reliability of  $\theta_Y$  from the second stage to the final stage for each of the designs in Table 6. The results showed that when finishing the first subdomain (i.e., the second stage finished), the longer the test length, the lower the RMSE and the higher the reliability in the second trait estimation, because the EAP brought information from the first trait estimation that became more precise as test lengths increased. After the third stage, panel 1-3-3-3 performed better than panels 1-3-2-3 and 1-3-1-3 because panel 1-3-3-3 had the largest bank size in the third stage. This matches Zenisky et al.'s (2009) finding that large bank sizes can improve measurement precision in MSTs. When all stages were completed, the differences between panels 1-3-3-3, 1-3-2-3, and 1-3-1-3 were reduced because the long test lengths made the measurements precise, regardless of the test design.

Table 3  
Mean and Standard Deviation of Bias for Two Latent Traits with Different Correlations in Distinct Panels

	Dimension 1			Dimension 2		
	$\rho_{XY} = .3$	$\rho_{XY} = .5$	$\rho_{XY} = .7$	$\rho_{XY} = .3$	$\rho_{XY} = .5$	$\rho_{XY} = .7$
Test length of each dimension = 20						
Panel 1-3-1-3	.000 (.004)	.002 (.005)	.000 (.004)	.000 (.004)	.001 (.005)	.000 (.004)
Panel 1-3-2-3	.000 (.004)	.002 (.005)	.000 (.004)	.000 (.004)	.001 (.005)	.001 (.004)
Panel 1-3-3-3	.000 (.004)	.002 (.005)	.000 (.004)	.000 (.004)	.001 (.005)	.001 (.004)
Test length of each dimension = 30						
Panel 1-3-1-3	.000 (.004)	.000 (.004)	.000 (.003)	.000 (.003)	-.001 (.004)	.000 (.003)
Panel 1-3-2-3	.000 (.004)	.000 (.004)	.000 (.003)	.000 (.004)	-.001 (.003)	.001 (.004)
Panel 1-3-3-3	.000 (.004)	-.001 (.004)	.000 (.003)	.001 (.003)	-.001 (.004)	.001 (.003)
Test length of each dimension = 40						
Panel 1-3-1-3	-.001 (.003)	-.001 (.003)	.000 (.003)	.000 (.003)	.000 (.003)	.000 (.003)
Panel 1-3-2-3	-.001 (.003)	-.001 (.003)	.000 (.003)	.000 (.003)	.001 (.003)	.000 (.003)
Panel 1-3-3-3	-.001 (.003)	-.001 (.003)	.000 (.003)	.000 (.003)	.000 (.003)	.000 (.003)

Note. The values inside the brackets are the standard deviations.

Table 4  
Mean and Standard Deviation of RMSE for Two Latent Traits with Different Correlations in Distinct Panels

	Dimension 1			Dimension 2		
	$\rho_{XY} = .3$	$\rho_{XY} = .5$	$\rho_{XY} = .7$	$\rho_{XY} = .3$	$\rho_{XY} = .5$	$\rho_{XY} = .7$
Test length of each dimension = 20						
Panel 1-3-1-3	.444 (.013)	.436 (.014)	.420 (.011)	.441 (.014)	.434 (.013)	.421 (.009)
Panel 1-3-2-3	.444 (.013)	.436 (.014)	.420 (.011)	.441 (.010)	.429 (.010)	.413 (.008)
Panel 1-3-3-3	.444 (.013)	.436 (.014)	.419 (.011)	.431 (.009)	.425 (.009)	.408 (.007)
Test length of each dimension = 30						
Panel 1-3-1-3	.374 (.009)	.369 (.008)	.358 (.008)	.375 (.009)	.369 (.008)	.358 (.008)
Panel 1-3-2-3	.374 (.009)	.369 (.008)	.358 (.008)	.376 (.008)	.367 (.008)	.353 (.008)
Panel 1-3-3-3	.374 (.009)	.369 (.008)	.358 (.008)	.367 (.008)	.362 (.007)	.348 (.007)
Test length of each dimension = 40						
Panel 1-3-1-3	.331 (.007)	.327 (.007)	.317 (.006)	.330 (.008)	.326 (.007)	.318 (.008)
Panel 1-3-2-3	.331 (.007)	.327 (.007)	.317 (.006)	.330 (.007)	.323 (.006)	.313 (.006)
Panel 1-3-3-3	.331 (.007)	.327 (.007)	.317 (.006)	.323 (.006)	.318 (.006)	.308 (.006)

Note. The values inside the brackets are the standard deviations.

Table 5  
*Reliability and Standard Errors of Reliability for Two Latent Traits with Different Correlations in Distinct Panels*

	Dimension 1			Dimension 2		
	$\rho_{XY} = .3$	$\rho_{XY} = .5$	$\rho_{XY} = .7$	$\rho_{XY} = .3$	$\rho_{XY} = .5$	$\rho_{XY} = .7$
Test length of each dimension = 20						
Panel 1-3-1-3	.802 (.011)	.810 (.012)	.824 (.009)	.806 (.012)	.811 (.011)	.823 (.007)
Panel 1-3-2-3	.802 (.011)	.810 (.012)	.824 (.009)	.805 (.009)	.815 (.009)	.829 (.006)
Panel 1-3-3-3	.802 (.011)	.810 (.012)	.824 (.009)	.814 (.008)	.819 (.008)	.833 (.006)
Test length of each dimension = 30						
Panel 1-3-1-3	.861 (.007)	.864 (.006)	.872 (.006)	.859 (.007)	.865 (.006)	.872 (.006)
Panel 1-3-2-3	.861 (.007)	.864 (.006)	.872 (.006)	.859 (.006)	.866 (.006)	.875 (.005)
Panel 1-3-3-3	.861 (.007)	.864 (.006)	.872 (.006)	.865 (.006)	.870 (.005)	.879 (.005)
Test length of each dimension = 40						
Panel 1-3-1-3	.890 (.005)	.893 (.004)	.899 (.004)	.892 (.005)	.893 (.005)	.898 (.005)
Panel 1-3-2-3	.890 (.005)	.893 (.004)	.899 (.004)	.892 (.004)	.896 (.004)	.902 (.004)
Panel 1-3-3-3	.890 (.005)	.893 (.004)	.899 (.004)	.896 (.004)	.899 (.004)	.905 (.004)

*Note.* The values inside the brackets are the standard errors.

Table 6  
*RMSE and Reliability for the Second Latent Trait in finishing Second, Third, and Final Stages When Correlation between Traits of 7*

	RMSE			Reliability		
	2nd Stage	3rd Stage	Final Stage	2nd Stage	3rd Stage	Final Stage
Test length of each dimension = 20						
Panel 1-3-1-3	-	.542 (.015)	.420 (.011)	-	.707 (.015)	.823 (.007)
Panel 1-3-2-3	.779 (.006)	.526 (.014)	.413 (.008)	.393 (.007)	.724 (.014)	.829 (.006)
Panel 1-3-3-3	.779 (.006)	.516 (.012)	.408 (.007)	.393 (.007)	.733 (.012)	.833 (.006)
Test length of each dimension = 30						
Panel 1-3-1-3	-	.475 (.014)	.358 (.008)	-	.774 (.014)	.872 (.006)
Panel 1-3-2-3	.762 (.006)	.463 (.012)	.353 (.008)	.421 (.004)	.785 (.010)	.875 (.005)
Panel 1-3-3-3	.762 (.006)	0.452 (.011)	.348 (.007)	.421 (.004)	.796 (.010)	.879 (.005)
Test length of each dimension = 40						
Panel 1-3-1-3	-	.432 (.014)	.317 (.006)	-	.813 (.014)	.898 (.005)
Panel 1-3-2-3	.751 (.006)	.419 (.008)	.313 (.006)	.436 (.004)	.824 (.010)	.902 (0.004)
Panel 1-3-3-3	.751 (.006)	.408 (.009)	.308 (.006)	.436 (.004)	.833 (.010)	.905 (.004)

*Note.* The values inside the brackets are the standard deviations for the RMSE and the standard errors for reliability. Panels 1-3-1-3 do not follow the ISR design; therefore, there are no estimates of the second latent trait at the end of the second stage.

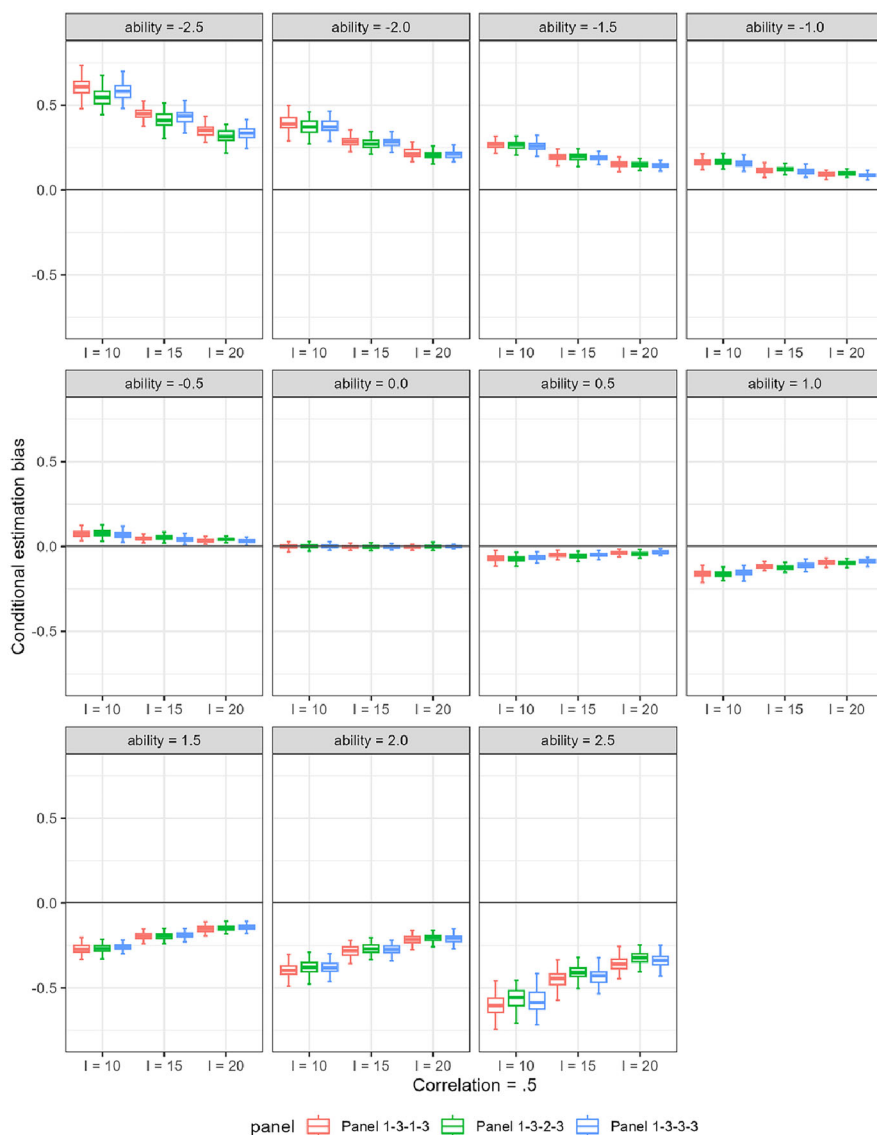


Figure 5. Conditional estimation bias in latent trait  $Y$  when the correlation between traits of 5. Note.  $I$  = the size of modules.

### Empirical Example Data Analysis

To generalize the results of the modified ISR design and demonstrate its applicability to operational tests, we used an empirical item bank and real-world abilities of 11,112 students. We compared the performance of the ISR design with that of the multiple unidimensional MST design in terms of bias, RMSE, and reliability. The data were obtained from the Test of Chinese as a Foreign Language (TOCFL)

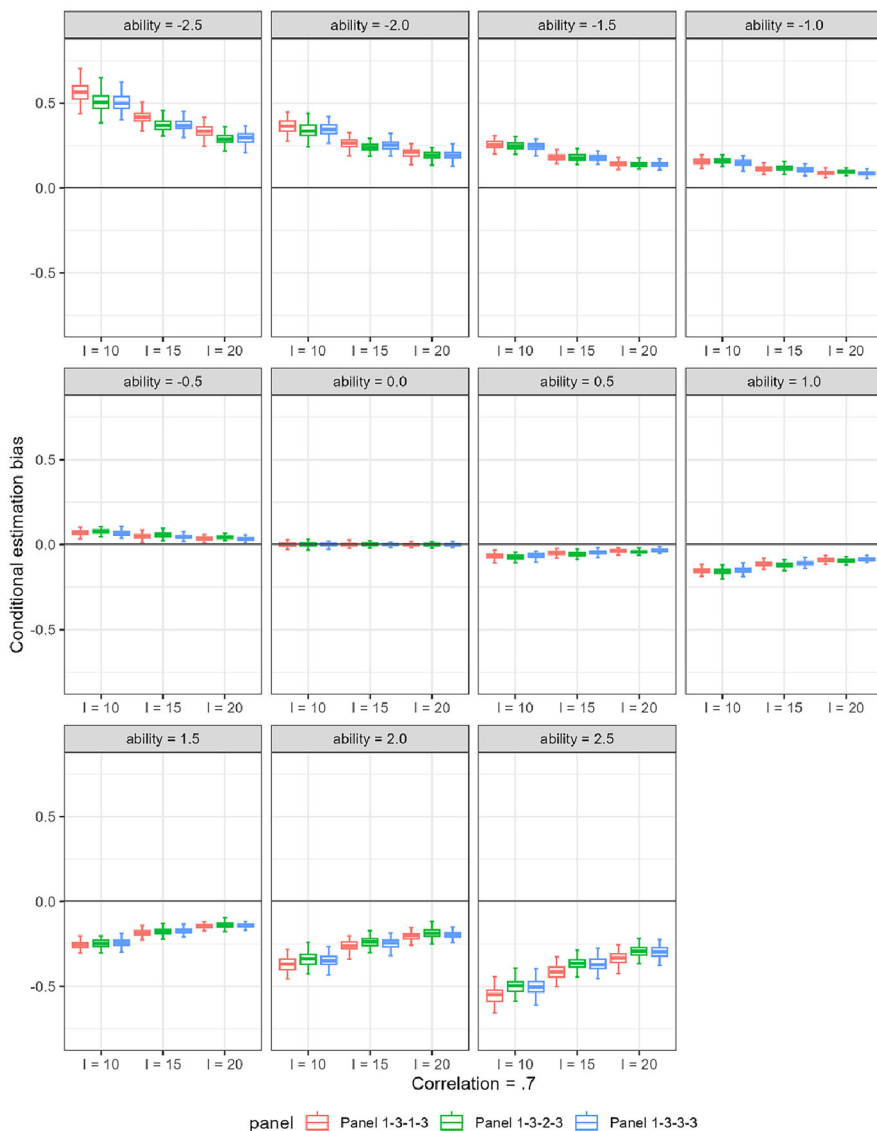


Figure 6. Conditional estimation bias in latent trait  $Y$  when the correlation between traits of 7. Note.  $I$  = the size of modules.

Speedy Screening Test, which comprises 110 multiple-choice items selected from an item bank. The test was administered to nonnative speakers in Taiwan in 2023. The Speedy Screening Listening and Reading tests developed by the Steering Committee for the Test of Proficiency (Huayu) adopted an MST framework, as shown in the top plot of Figure 7. The examinees took a listening test before the reading test. Each test adopted a 1-3-3 MST design. For both the listening and reading tests, all discrimination parameters in the item bank were 1.0 for both the listening and reading tests.

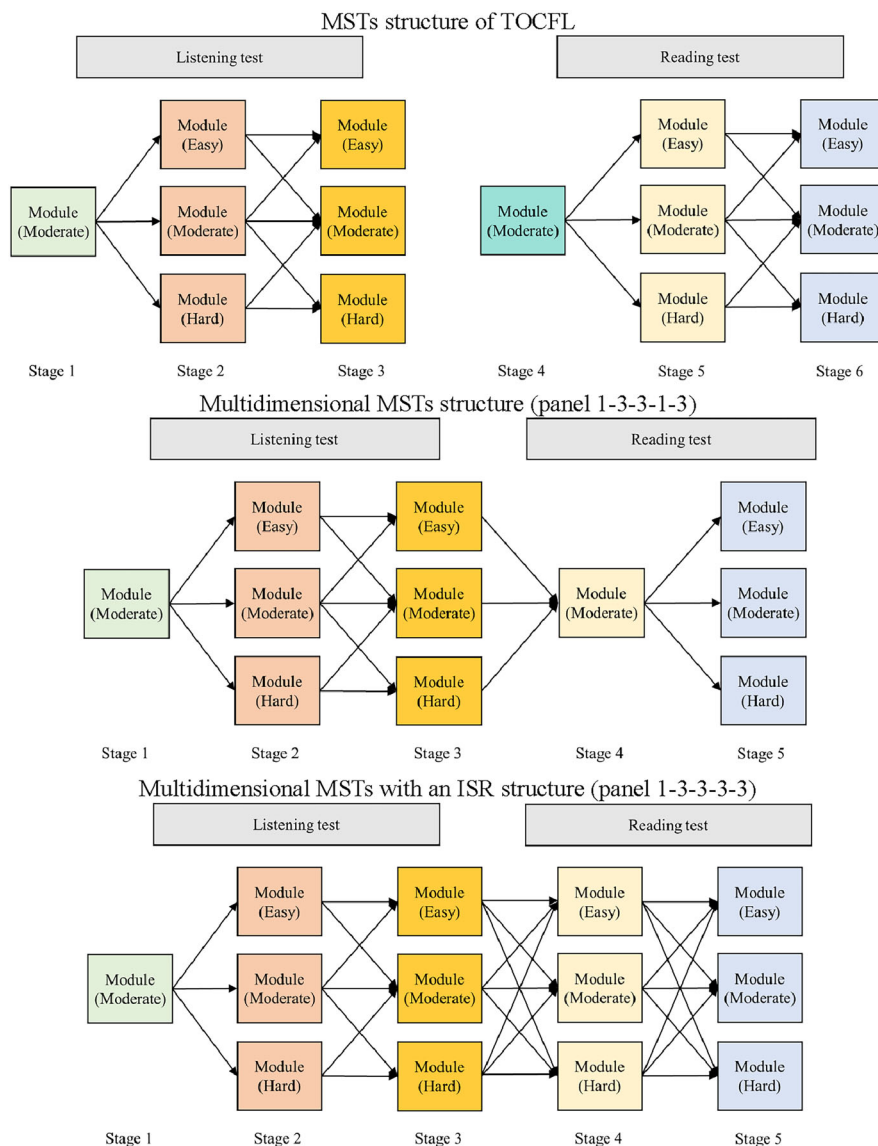


Figure 7. MSTs structure of TOCFL and empirical data.

The difficulty parameters in the listening test ranged from  $-8.04$  to  $2.99$ , with a mean of  $2.00$  and a standard deviation of  $2.53$ . In the reading test, the difficulty parameters ranged from  $-8.26$  to  $3.66$ , with a mean of  $1.51$  and a standard deviation of  $2.86$ . The two traits followed a multivariate normal distribution, with a mean of  $[-.92, -1.17]$  and a correlation of  $.82$ .

The item and person parameters calibrated by fitting the 2PL model to real-world response data were adopted as true parameters in this empirical study. The real-world

test (TOCFL) in the empirical analysis was a high-stakes test designed using a traditional MST without an ISR design. This means that we can only observe the performance of non-ISR designs, not ISR designs because we do not have empirical data working under ISR designs. To make a comparison between ISR design and non-ISR design, we copied the item parameters and MST design obtained from the real-world data and used those parameters and test design to generate responses in the two comparable designs.

We compared the two designs, panels 1-3-3-1-3 (traditional MST) and 1-3-3-3-3 (MST with ISR), as shown in the middle and bottom plots in Figure 7, where the first three stages measured listening proficiency and the last two stages measured reading proficiency. The responses to the listening test items (Stages 1, 2, and 3) and the corresponding rootings were fixed to the real-world responses; however, the responses to the reading test items (Stages 4 and 5) were generated from the 2PL model with the true values of the item and person parameters. The responses and module selections at Stages 4 and 5 were simulated with 100 replicates. This is because the original TOCFL MST did not have an ISR design for any of the tests. In the first dimension, the listening test contained 10 items in Stage 1, 8 items in Stage 2, and 7 items in Stage 3, for a total of 25 items. The reading test contained 8 items in Stage 4 and 7 items in Stage 5, for a total of 15 items.

We reported the bias, RMSE, and reliability only for the reading test but not for the listening test because the responses to the listening test items were fixed across replicates. The results showed that the bias of reading ability estimation had a mean of .061 and a standard deviation of .003 for panel 1-3-3-1-3 and a mean of .059 and a standard deviation of .003 for panel 1-3-3-3-3. Both designs yielded a slightly positive bias in the estimation of reading proficiency. This bias can be attributed to the mean item difficulty of 1.51, which exceeds the mean ability at the population level. This finding aligns with Lord's (1986) report on the EAP estimator's tendency to overestimate low ability levels. The RMSE for panel 1-3-3-1-3 had a mean of .461 and a standard deviation of .003, whereas that for panel 1-3-3-3-3 had a mean of .451 and a standard deviation of .003. Panel 1-3-3-3-3 reduced the RMSE and improved the measurement precision compared to panel 1-3-3-1-3 because Stage 4 allowed module selection by considering the intertrait correlation and information from the listening test in panel 1-3-3-3-3. The reliability of the reading test supported this conclusion. Reliability had a mean of .889 and a standard deviation smaller than .001 in panels 1-3-3-1-3, while a mean of .894 and a standard deviation smaller than .001 in panel 1-3-3-3-3. All the results of the empirical studies were consistent with the simulation studies in that the ISR design improved the measurement precision of the second latent trait, but the improvement was limited.

## Discussion

In this study, we extended Han's framework to a multidimensional test by utilizing the correlation coefficients between latent traits to effectively establish an item bank through MIRT modeling. However, the limitations of this approach must be acknowledged when using the MAP or EAP methods. These methods are known to result in estimates biased toward the center of a prior distribution, resulting in a

shrunken score scale (Weiss & McBride, 1984). Consequently, the estimation of  $\theta$  was underestimated when  $\theta > 0$  and overestimated when  $\theta < 0$ . This finding is consistent with the property of Bayesian-based estimation methods, known as shrinkage of scale (Baker & Kim, 2004; McBride, 1977; Novick & Jackson, 1974; Weiss & McBride, 1984).

The implications of RMSE and reliability patterns in the results of this study are as follows. First, the higher the intertraits correlation, the better the improvement of RMSE and reliability by using ISR design compared to non-ISR design. The reliability can be improved by .01 in the .7 intertrait correlation condition, whereas by .008 in the .3 intertrait correlation condition. It implied that ISR design might benefit better for the target participants with high correlations between traits. Especially for children or adolescents, literature has discovered high intertrait correlations between cognitive abilities. For example, the correlation between reading and mathematics scores was around .60 for primary school students and adolescents (Hecht et al., 2001) and even higher for preschool children (McClelland et al., 2007). Second, the advantage of the ISR design was greater in the short test length condition than in the long test length condition. The ISR design had .026 better reliability than the non-ISR design in a 10-item test and .01 better in a 20-item test, whereas .006 better in a 40-item test when the intertrait correlation is equal to .7. It implied that when the test length is strictly limited in practice, using the ISR design would be a helpful approach to improve the test reliability. Especially for the test program measuring a large number of correlated traits within a limited test time, it is better to shorten the test lengths for each trait to avoid the fatigue effect. In such a case, using the ISR design for the short-length MST is highly recommended for better reliability than non-ISR design.

We provide practical guidelines for applying ISR in MST based on our results here. First, when using Equation 2 to predict the current trait level, utilizing high-correlated variables (covariates) is better for the ISR design. We recommend incorporating as many correlated traits as possible and even demographic variables together when using the ISR design with the covariance matrix between all correlated covariates (latent or observed) in Equation 2. This is because the posterior variance  $\sigma_{h|c}^2 = \sigma_h^2 - \Sigma_{hc} \Sigma_{cc}^{-1} \Sigma'_{hc}$  is smaller (i.e., a higher precision in predicting the current trait) when  $\Sigma_{hc} \Sigma_{cc}^{-1} \Sigma'_{hc}$  is larger, which can be achieved by adding more variables to the c vector. It is reasonable and similar to the concepts in regression, where more predictors make the unexplained residual variance smaller. When we include a variable with a correlation above .7 to predict the current trait, we expect ISR design to improve by .024 of test reliability or more compared to non-ISR design in the short test length, such as 10 items. Second, we can establish an ISR design from a non-ISR design without requiring new items in the item bank by shortening the test length. When we plan to shorten the test length of existing MSTs, the existing item bank for a long test length situation allows us to assemble more test modules for a short test length situation in MST, which is sufficient to implement ISR design. We acknowledged that ISR design requires more test modules in the item bank than non-ISR design. For example, in our simulation study, the 1-3-3-3 design (ISR) requires two more test modules than the 1-3-1-3 design (non-ISR). We can assemble more

test modules with the same item bank size when we shorten the test length in MST. This means we do not need to develop new items when we shorten the test length to establish an ISR design that contributes significantly to short test length reliability.

We recognize that the conditional bias and RMSEs did not change significantly from the unidimensional MST to ISR designs (Figures 5 and 6). The test length had a more significant effect on measurement precision than the ISR design. Although the improvement in measurement precision was limited, slight additional effort was required for test designers to implement the ISR design in practice. The effort test designers should change unidimensional MST to ISR designs: (1) update the ability estimator from unidimensional to multidimensional EAP and (2) allow the first stage of subsequent subtests to have multiple modules that would be adaptively selected according to ability estimates calculated from Equation 2. Multiple modules can be assembled automatically using the automated test assembly proposed by van der Linden (2005) and Diao and van der Linden (2011). This is not a difficult change for test designers when both the item bank and MST administration systems have been established.

From Tables 3–5, the bias, RMSE, and reliability of trait 1 remained unchanged across ISR designs because trait 1 was not estimated at later stages after the domain test measuring it. Theoretically, estimating trait 1 after completing all stages in ISR design could improve its precision. However, in this study, we did not re-estimate trait 1 after its domain test. Even if we had, the expected improvement in precision would likely be minimal and negligible, as evidenced by the limited improvement observed for trait 2. Given the slight enhancement in trait 2 precision, any indirect effect on trait 1 estimates through prior correlation would be even smaller. For instance, in Table 4, under the condition of  $\rho_{XY} = .7$  and a test length of 20, the RMSE of trait 2 improved marginally from .421 in the 1-3-1-3 design to .408 in the 1-3-3-3 design. This slight improvement can be attributed entirely to the differences in item selection design. When trait 1 is estimated at later stages during the test for trait 2, precision improvement for trait 1 is driven only by the prior intertrait correlation in the EAP estimator. However, with a test length of 20 items, the likelihood provides substantially more information than the prior, rendering the contribution of prior correlation to test information negligible.

Another practical issue is that the ISR design requires a larger sample size to update the item parameters in the item bank than a non-ISR design. Compared to the non-ISR design, the ISR design contains a larger bank size in the first stages of every subtest because it allows multiple modules at such stages; however, the non-ISR design does not allow this. Conditional on the same number of examinees, a larger bank leads to fewer responses per item, which may cause larger standard errors in item parameter estimation when updating the item parameters for MST in practice. This is a common issue in adaptive test designs. Larger banks require larger sample sizes for updated item calibration. Allowing a larger bank size at the first stages of specific dimensions than the non-ISR design is also one of the advantages of the ISR design because the ISR design can easily become a non-ISR design if we do not have a sufficient sample size for updating item parameters.

However, in this study, we did not consider the potential measurement errors or biases of the item parameters. The bias in item parameter estimates results in bias in

proficiency estimation, and the bias for examinees relies on how they are routed (Wu & Xi, 2017). Jewsbury and van Rijn (2020) showed that this bias is caused by the difference between multiple IRT and MIRT in the handling of missing data, which is inherent in adaptive testing. This study did not examine the effects of estimating the item parameters. In practice, all item parameters in the item bank must be calibrated. Especially in the case of small sample sizes, bias in estimating item parameters could increase the uncertainty in the estimation of examinees' abilities (Patton et al., 2013; van der Linden & Glas, 2000). Therefore, in this study, all situations were simulated under the assumption of known and correct parameters.

Finally, by utilizing empirical data from the TOCFL and our multidimensional MST with an ISR design, we conducted a comprehensive application to operational testing. The results indicate that compared to the traditional unidimensional MST design, our proposed ISR design demonstrates favorable performance in terms of bias, RMSE, and reliability. This study underscores the potential of ISR design to enhance test efficiency and accuracy, and provides valuable insights for future advancements in test design and assessment methods.

Future studies may consider a more complex MIRT model to control for undesirable confounding effects, such as the item position effect (Debeer & Janssen, 2013), which is related to examinees' efforts in LSA. The effort effect of examinees can be captured as a nuisance factor. Under Veldkamp and van der Linden's (2002) MCAT framework, the current study belongs only to the case of a simple multidimensional structure in which all abilities are intentional. Furthermore, they introduced a case in which the nuisance and intentional dimensions existed concurrently during the MCATs. The ISR design that considers nuisance factors across subdomains and stages in multidimensional MSTs has room for discussion in future studies.

Additionally, this study specifically considers between-item multidimensional tests (Adams et al., 1997), but in practical scenarios, items may measure multiple abilities within themselves. Consequently, the MST framework has the potential to be extended to encompass within-item multidimensional testing. While this study explored various conditions, including different ISR designs within the MST, the relationships between latent traits, and the number of items in each module, it is important to recognize that the findings may not be universally applicable to all MST conditions and designs, as these factors can significantly impact MST performance.

### **Acknowledgments**

This study was supported by the National Taiwan Normal University (NTNU), Taiwan, ROC, and Professor Greg Lee from the Department of Computer Science and Information Engineering at NTNU. We sincerely appreciate the financial support from NTNU and Professor Lee for English language editing services. Additionally, we thank the Steering Committee for the Test of Proficiency-Huayu for agreeing to release the Test of Chinese as a Foreign Language data used in this study. The responsibility for errors in fact or judgment is ours.

## Appendix: Pseudo Code for the Simulation of MST with ISR

---

**Input:**

Parameters of item bank: Q

Response of examinees: R

Module: M

Multivariate prior distribution of traits: H

Estimated abilities for completed dimensions: ETC

**Output:**

Estimation ability for current dimension: ET

ET ← Equation\_2(H, ETC)

**if** at the first stage of MST **then**

    max(M) ← maximum information method(ET, Q)

    ET ← EAP method(Q, R, H)

**for** stage 2 **to** stage 4 **do**

    max(M) ← maximum information method(ET, Q)

    ET ← EAP method(Q, R, H)

**return** ET

---

## References

- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. <https://doi.org/10.1177/0146621697211001>
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Basel.
- 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>
- Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*(1), 5–20. <https://doi.org/10.1177/0013164406288162>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1–39. <https://doi.org/10.18637/jss.v071.i05>
- Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164–185. <https://doi.org/10.1111/jedm.12009>
- Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp\_solve version 5.5 in R. *Applied Psychological Measurement, 35*(5), 398–409. <https://doi.org/10.1177/0146621610392211>
- Ding, H., & Homer, M. (2020). Interpreting mathematics performance in PISA: Taking account of reading performance. *International Journal of Educational Research, 102*, 101566. <https://doi.org/10.1016/j.ijer.2020.101566>

- Eaton, M. L. (1983). *Multivariate statistics: A vector space approach*. Wiley. <https://doi.org/10.2307/2347710>
- Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19*, 221–239. [https://doi.org/10.1207/s15324818ame1903\\_4](https://doi.org/10.1207/s15324818ame1903_4)
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement, 37*, 259–275. <https://doi.org/10.1177/0146621612473638>
- Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement, 40*(4), 289–301. <https://doi.org/10.1177/0146621616631317>
- Han, K. T. (2020). Framework for developing multistage testing with intersectional routing for short-length tests. *Applied Psychological Measurement, 44*, 87–102. <https://doi.org/10.1177/0146621619837226>
- Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology, 79*(2), 192–227. <https://doi.org/10.1006/jecp.2000.2586>
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*, 44–152. <https://doi.org/10.1111/j.1745-3992.2007.00093.x>
- 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*, 383–402. <https://doi.org/10.3102/1076998619881790>
- Khorrandel, L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2020). Plausible values: Principles of item response theory and multiple imputations. In D. B. Maehler, & B. Rammstedt (Eds.), *Large-scale cognitive assessment: Analyzing PIAAC data* (pp. 27–48). Springer. <https://doi.org/10.1007/978-3-030-47515-4>
- Lord, F. M., Novick, M. R. & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge. <https://doi.org/10.4324/9780203056615>
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157–162. <https://www.jstor.org/stable/1434513>
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229–249. <https://doi.org/10.1111/j.1745-3984.1998.tb00537.x>
- Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189–202. [https://doi.org/10.1207/s15324818ame1903\\_2](https://doi.org/10.1207/s15324818ame1903_2)
- McBride, J. R. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*, 121–140. <https://doi.org/10.1177/014662167700100119>
- McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary,

- and math skills. *Developmental psychology*, 43(4), 947–959. <https://doi.org/10.1037/0012-1649.43.4.947>
- McKinley, R. L., & Reckase, M. D. (1982). The investigation of two-stage adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 6, 181–197. <https://doi.org/10.1177/014662168200600206>
- Min, S., & Bishop, K. (2024). A shortened test is feasible: Evaluating a large-scale multistage adaptive English language assessment. *Language Testing*, 41(3), 627–648. <https://doi.org/10.1177/02655322231225426>
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics*, 17, 131–154. <https://doi.org/10.3102/10769986017002131>
- Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. McGraw-Hill.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351–356. <https://doi.org/10.2307/2285821>
- Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Unpublished doctoral dissertation). University of Massachusetts Amherst. Retrieved from: <https://www.proquest.com/openview/b35c8cc5459c2cf54469edf5d3eaaf1fb/1>
- Patsula L. N., & Hambleton R. K. (1999, April). *A comparative study of ability estimation from computer-adaptive testing and multistage testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec.
- Patton, J. M., Cheng, Y., Yuan, K. H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, 37, 24–40. <https://doi.org/10.1177/0146621612461727>
- Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin*, 146(7), 595. <https://doi.org/10.1037/bul0000231>
- R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R foundation for statistical computing. <https://www.R-project.org/>.
- Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multistage adaptive testlet design*. Computerized Testing Report 97-02. Newtown, PA: Law School Admissions Council.
- Reckase, M. D. (2009). Computerized adaptive testing using MIRT. In M. D. Reckase (Eds.), *Multidimensional item response theory* (pp. 311–339). Springer.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331–354. <https://doi.org/10.1007/BF02294343>
- Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement*, 21, 129–142. <https://doi.org/10.1177/01466216970212003>
- Tang, X., Zheng, Y., Wu, T., Hau, K. T., & Chang, H. H. (2024). Utilizing response time for item selection in on-the-fly multistage adaptive testing for PISA assessment. *Journal of Educational Measurement*. <https://doi.org/10.1111/jedm.12403>
- Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, 16(1), 1–9. <https://doi.org/10.7275/wqzt-9427>

- Ünal, Z. E., Greene, N. R., Lin, X., & Geary, D. C. (2023). What is the source of the correlation between reading and mathematics achievement? Two meta-analytic studies. *Educational Psychology Review*, *35*(1), 4. <https://doi.org/10.1007/s10648-023-09717-5>
- van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer.
- van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*, 35–53. [https://doi.org/10.1207/s15324818ame1301\\_2](https://doi.org/10.1207/s15324818ame1301_2)
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of computerized adaptive testing*. Springer.
- van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer.
- Vispoel, W. P. (1998). Reviewing and changing answers on computer adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*, 328–345. <https://doi.org/10.1111/j.1745-3984.1998.tb00542.x>
- Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, *37*, 21–38. <https://doi.org/10.1111/j.1745-3984.2000.tb01074.x>
- Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588. <https://doi.org/10.1007/BF02295132>
- von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 1039–1055). Amsterdam, the Netherlands: Elsevier.
- Wainer, H. (1992). Some practical considerations when converting a linearly administered test to an adaptive format. *ETS Research Report Series*, *1992*(1), i–11. <https://doi.org/10.1002/j.2333-8504.1992.tb01445.x>
- Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates Publishers.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. <https://doi.org/10.1007/BF02294627>
- Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 295–316. <https://doi.org/10.1177/0146621604265938>
- Wang, S., Fellouris, G., & Chang, H. H. (2017). Computerized adaptive testing that allows for response revision: design and asymptotic theory. *Statistica Sinica*, *27*, 1987–2010. <https://doi.org/10.5705/ss.202015.0304>
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
- Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement*, *8*, 273–285. <https://doi.org/10.1177/014662168400800303>
- Weissman, A. (2014). IRT-based multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 153–168). Boca Raton, FL: Chapman and Hall/CRC.
- Wu, M., & Xi, N. (2017). *Multistage testing in the 2015 NAEP mathematics DBA field trial*. Paper presented at the NCME Annual Meeting, San Antonio, TX.

- Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*, 347–368.
- Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*, 16–27. <https://doi.org/10.1111/emip.12226>
- Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Introduction of multistage adaptive testing design in PISA 2018. OECD Education Working Papers*, 209. Paris: OECD Publishing. <https://doi.org/10.1787/b9435d4b-en>.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>
- Zenisky, A. L. (2004). *Evaluating the effects of several multistage testing design variables on selected psychometric outcomes for certification and licensure assessment*, Unpublished doctoral dissertation. University of Massachusetts, Amherst.
- Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2009). Multistage testing: Issues, designs and research. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer. [https://doi.org/10.1007/978-0-387-85461-8\\_18](https://doi.org/10.1007/978-0-387-85461-8_18)

### Authors

- Yi-Ling Wu is a Postdoctoral Fellow at the Department of Computer Science and Information Engineering, National Taiwan Normal University, No.88, Sec. 4, Tingzhou Rd., Wenshan Dist., Taipei City 116, Taiwan; [ylwu@csie.ntnu.edu.tw](mailto:ylwu@csie.ntnu.edu.tw). Her primary research interests include item response theory models, multistage adaptive testing, and statistical methods.
- Yao-Hsuan Huang is a PhD student at Department of Educational Psychology and Counseling, National Taiwan Normal University, No. 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan; [randoll9th@gmail.com](mailto:randoll9th@gmail.com). His primary research interests include item response theory models and computerized adaptive testing.
- Chia-Wen Chen is a Psychometrician at the Psychometrics Centre, Cambridge Judge Business School, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, United Kingdom; [c.chen@jbs.cam.ac.uk](mailto:c.chen@jbs.cam.ac.uk). His primary research interests include item response theory models, multilevel models, mixture models, computerized adaptive testing, forced-choice items, and differential item functioning.
- Po-Hsi Chen is a Professor at the Department of Educational Psychology and Counseling, Director of Research Center for Psychological and Educational Testing, and a Professor of the Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, No. 162, Sec. 1, Heping E. Rd., Da'an Dist., Taipei City 106209, Taiwan; [chenph@ntnu.edu.tw](mailto:chenph@ntnu.edu.tw). His primary research interests include item response theory models and multidimensional computerized adaptive testing.