frontiers in Psychology

ORIGINAL RESEARCH
published: 05 March 2019
doi: 10.3389/fpsyg.2019.00240

Item Selection Methods for Computer Adaptive Testing With Passages

*Lihua Yao**

Office of People Analytics, Defense Personnel Assessment Center, Defense Human Resource Activity, United States Department of Defense, Seaside, CA, United States

Computer adaptive testing (CAT) has been shown to shorten the test length and increase the precision of latent trait estimates. Oftentimes, test takers are asked to respond to several items that are related to the same passage. The purpose of this study is to explore three CAT item selection techniques for items of the same passages and to provide recommendations and guidance for item selection methods that yield better latent trait estimates. Using simulation, the study compared three models in CAT item selection with passages: (a) the testlet-effect model (T); (b) the passage model (P); and (c) the unidimensional IRT model (U). For the T model, the bifactor model with testlet-effect or constrained multidimensional IRT model was applied. For each of the three models, three procedures were applied: (a) no item exposure control; (b) item exposure control of rate 0.2 ; and (c) item exposure control of rate 1. It was found that the testlet-effect model performed better than passage or unidimensional models. The P and U models tended to overestimate the precision of the theta or latent trait estimates.

Keywords: CAT, item response theory, multidimensional item response theory, MLE, IRT, MIRT, testlet-effect, passages

OPEN ACCESS

Edited by:
Weijun Wang,
University at Buffalo, United States

Reviewed by:
Min Wang,
Texas Tech University, United States
Haiming Zhou,
Northern Illinois University, United States

***Correspondence:**
Lihua Yao
lihua.yao.civ@mail.mil

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *05 October 2018*
Accepted: *24 January 2019*
Published: *05 March 2019*

Citation:
Yao L (2019) Item Selection Methods for Computer Adaptive Testing With Passages. Front. Psychol. 10:240. doi: 10.3389/fpsyg.2019.00240

INTRODUCTION

In Reading, Mathematics, Listening, and English tests, students are often asked to respond to several items related to a common stimulus. Information used to answer these items is interrelated in the passage. These kinds of assessments are known to be likely to produce local item dependence (LID). Oftentimes, for simplicity, the related items are scored independently or are summed as one score. Yen (1993) points out two negative measurement effects of ignoring LID in standard item response theory (IRT) parameter estimation and scoring, namely overestimation of the precision of prociency estimates and bias in discrimination parameter estimates.

A testlet can be created by combining the set of inter-related items. For dichotomous scored items, testlet effect has been modeled to include a random testlet effect by modifying the two parameter model (testlet-effect-2PL; Bradlow et al., 1999) and the three-parameter model (testlet-effect-3PL; Waniner et al., 2000), Rasch model (testlet-effect-Rasch; Wang and Wilson, 2005). DeMars (2006) applied the Bi-Factor multidimensional three-parameter logistic (M-3PL) IRT model to the testlet-based test. The testlet-effect-3Pl model is a constrained M-3Pl model in parameter estimation and ability estimation. These models were applied to tests of traditional paper and pencil format.

Computer adaptive testing (CAT) has the advantages of increasing the measurement precision; after each item selection, the examinee's ability is updated and the next selected item will maximally improve the accuracy for this estimated examinee. Testing companies administrate items of passages for content areas such as Reading and Paragraph Comprehension. There are usually two methods in CAT item selection. One method is to use the unidimensional IRT (UIRT) CAT item selection method and treat each item as independent; even though the items in the passage are dependent. The other method is to select the passage with the maximum information or the minimum error variance and to have the items in the passages be all administrated or partially administrated; the information for the passage is the sum of the informations for all the items in the passage.

Murphy et al. (2010) studied and compared the 3PL IRT and 3PL TRT (testlet response theory) item selection methods, and found that they performed similarly. However, the information for the passage is computed based on the sum of item information under the 3PL IRT model, and item dependency is ignored during item selection. Models taking into account the item dependency in item selection for CAT have not been studied.

This study used simulation and compared three models in CAT item selection: (a) the testlet-effect model (T); (b) the passage model(P); and (c) the unidimensional IRT model (U). The newly proposed T model for CAT is compared with two commonly used P and U models. For the T model, the bifactor model (M-3PL) with testlet-effect was applied. The information for each testlet was computed based on M-3Pl model and was used as the item selection criteria. For the P model, the information for the passage was computed based on the sum of item information under the 3PL IRT model, and item dependency was ignored during item selection; this is similar to the TRT model in Murphy et al. (2010). For the U model, the regular 3PL model was applied and the information for each item was used as the item selection criteria.

For each of the three models, three procedures were applied: (a) no item exposure control; (b) item exposure control of rate 0.2 using priority index method; and (c) item exposure control of rate 1 using priority index method. Priority Index (PI) is a method that puts probability or weight on each item in the pool after each item selection.

Various item selection criteria have been proposed and studied. Murphy et al. (2010) has applied MFI (maximum information), MPWI (maximum the posterior weighted information), and MEPV (minimum the expected posterior variance); they found no or little difference among these. For both MPWI and MEPV, integral or weighted summation would require much longer computation time for multidimensional models (T model here). Other methods such as Kullback-Leibler information (Chang and Ying, 1996; Veldkamp and van der Linden, 2002) and Volume (Segall, 1996) methods have been studied and compared in the MIRT frame work (Yao, 2012). The testlet-effect model in this paper was a constrained MIRT model with 39 dimensions. Thus, only the maximum information item selection criteria was applied, as other methods would require much longer computer time.

Multidimensional Models and TestLet-Effect Models

For a dichotomous scored or multiple choice item j, the probability of a correct response to item j for an examinee with ability $\vec{\theta}_i = (\theta_{i1}, \cdots, \theta_{iD})$, following the multidimensional three-parameter logistic (M-3PL; Reckase, 1997, 2009) model is defined as:

$$P_{ij1} = P(x_{ij} = 1 \mid \vec{\theta}_i, \vec{\beta}_j) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\vec{\beta}_{2j} \odot \vec{\theta}_i^T + \beta_{1j})}}, \quad (1)$$

where $x_{ij} = 0$ or 1 is the response of examinee i to item j. $\vec{\beta}_{2j} = (\beta_{2j1}, \cdots, \beta_{2jD})$ is a vector of dimension D for item discrimination parameters. β_{1j} is the scale difficulty parameter. β_{3j} is the scale guessing parameter. Here the dot product is defined as $\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$. The parameters for the jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j})$.

For a polytomous scored or a constructed response item j, the probability of a response $k-1$ to item j for an examinee with ability $\vec{\theta}_i$ is given by the multi-dimensional version of the partial credit model (M-2PPC; Yao and Schwarz, 2006) :

$$P_{ijk} = P(x_{ij} = k-1 \mid \vec{\theta}_i, \vec{\beta}_j) = \frac{e^{(k-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{k} \beta_{\delta tj}}}{\sum_{m=1}^{K_j} e^{((m-1)\vec{\beta}_{2j} \odot \vec{\theta}_i^T - \sum_{t=1}^{m} \beta_{\delta tj})}}, \quad (2)$$

where $x_{ij} = 0, \cdots, K_j - 1$ is the response of examinee i to item j. $\vec{\beta}_{2j} = (\beta_{2j1}, \cdots, \beta_{2jD})$ is a vector of dimension D for item discrimination parameters. $\beta_{\delta kj}$ for $k = 1, 2, \ldots, K_j$ are the threshold parameters or Alpha parameters, $\beta_{\delta 1j} = 0$, and K_j is the number of response categories for the jth item. The parameters for the jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{\delta 2j}, \ldots, \beta_{\delta K_j j})$.

Testlet-effect-2PPC/3PL model is a constrained M-2PPC/3PL model. This model essentially puts a constraint on the discrimination parameter within each testlet or cluster of inter-related items in a form of a constant. The discrimination parameter varies across testlets to account for the testlet effect. Suppose there are $D-1$ testlets for a test. Then the model can be D dimensional IRT model, and the discrimination parameters are

$$\vec{\beta}_{2j} = (\beta_{2j1}, \beta_{2j1}\gamma_1, \beta_{2j1}\gamma_2, \cdots, \beta_{2j1}\gamma_{D-1}) \quad (3)$$

where $\gamma = (\gamma_1, \cdots, \gamma_{D-1})$ are the variances of the testlet-effect parameters for the $D-1$ testlets. Within each testlet, the ratio of the item general discrimination (β_{2j1}) and the item testlet-effect discrimination is a constant, namely testlet-effect parameter γ_k, where $k \in \{1, \cdots, D-1\}$. The other item parameters (item difficulty/guessing or threshold) remain the same as the general MIRT model. For a testlet-effect model based on a common stimulus, each item belongs to only one testlet, i.e., the discrimination parameters for item j is $(\beta_{2j1}, \beta_{2j1}\gamma_{\delta_j})$, where $\delta_j \in \{1, 2, \cdots, D-1\}$. As in Li et al. (2006) or DeMars (2006), the formulas presented here are consistent with those found in the existing testlet models by Wainer et al. (2007). For item j in *kth* testlet-effect,

$$\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \beta_{2j1}\theta_{i1} + \beta_{2j1}\gamma_k\theta_{ik}, \quad (4)$$

and $\gamma_k\theta_{ik} \sim N(0, \gamma_k^2)$.

Unidimensional model is a special case of the multidimensional model where the dimension D is 1.

SIMULATION STUDY

The models introduced above were used for the simulation study. Item parameters introduced are unknown at the beginning; generally, a try out or a field test is conducted to collect responses from test-takers and the item parameters would then be estimated by an existing software. For CAT, item parameters in the pool should be known. Two set of item parameters are derived as described below using the real responses from test-takers taking paper and pencil format test.

Item Pool

Two sets of item pools were derived from BMIRT (Yao, 2003), based on real data listening assessments; there were 100 items with 38 testlets and 16 single items with 5000 examinees. Most of the testlets had two items, with some having 3 or 4 items. There were 16 items that were independent from each other and from others. The testlet-effect model of 39 dimensions and the unidimensional IRT models were applied to the data. The 38 estimated testlet-effect parameters varied from 0.05 to 0.7. Item Pool One had 100 testlte-effect item parameters. Item Pool Two had 100 UIRT item parameters. The two models fit the data equally well. The correlations between the estimates for the two models for the discrimination, difficulty, and guessing parameters were 0.97, 0.99, and 0.91, respectively; the first discrimination was used for the testlet-effect model. The AICs were 501917 and 527192 for the UIRT model and the testlet model, respectively. The chi-square difference between the two models was 1.78, which was considered as not significant.

Pool One had 100 items with item parameters of 39 dimensions following the testlet-effect model and was used for T model selection method. Pool Two had 100 items with item parameters of the unidimensional 3PL model and was used for U model selection. Pool Two was also used for the P model selection method.

The Priority Index for Item Exposure Control

The multidimensional priority index (MPI, Cheng and Chang, 2009; Yao, 2013) for each item *j* is defined by

$$MPI_j = \prod_{l=1}^{D} f_{jl}^{c_{jl}}, \tag{5}$$

where the constraint matrix $C = (c_{jl})_{J\times D}$, indicating the loading information for item *j* on domain *l*, is defined as the following:

$$c_{jl} = \begin{Bmatrix} 1 & \text{if item j load on domain l} \\ 0 & \text{otherwise} \end{Bmatrix}$$

For the *j*th item, let r_j denote its exposure rate. For each selection step, let n_j be the number of examinees that have already selected item *j*. The index for the item exposure control is defined by (van der Linden and Veldkamp, 2004, 2007; Yao, 2013)

$$f_{jl} = \max\{\frac{r_j - \frac{n_j}{N}}{r_j}, 0\}, \tag{6}$$

where *N* is the total number of examinees. This index will make sure that no item is selected with exposure rate larger than the predefined rate $\vec{r} = (r_1, \cdots, r_J)$. At the beginning of item selection, the weight or the probability of being selected for the item is high. If the item is administrated, then the weight or the probability for the item is smaller, until it reaches 0, then this item will not be selected anymore.

For the testlet-effect model and the passage selection model, there were only 54(38 + 16) passages or items in the pool. Thus, for 2,000 examinees, item exposure rate of 0.2 would result in no items being selected if all the item priority index were 0 after reaching exposure rate. Therefore, a modification was made after all items had reached exposure rate; the probability was reset for each item in the pool; although a higher exposure rate could be applied to avoid this problem.

Three methods were used for item exposure control in this study: no item exposure control and exposure rate of 0.2 and 1 using the Priority Index.

True Abilities

For this simulation, 2,000 examinees (true abilities) were sampled from the standard multivariate normal distribution of dimension 39; these examinees were used for the T model. For the P and the U model, the first/general ability was used.

CAT Item Selection Methods

For all item selection methods, the first item or passage was randomly selected from the top 20, and the second item or passage was randomly selected from the top 10. There were three types of models for the item selection methods: the testlet-effect model, the unidimensional passage based model, and the unidimensional IRT based model. Maximum information was used as the criteria for all methods. MAP ability estimates were used to update ability after each item or passage selection. Testlet lengths of 10, 20, and 30 were specified. For the testlet-effect model and the passage selection models, if all items in the selected testlet or passage were administered, then the actual number of items selected for each examine might have been slightly higher than 10, 20 or 30, because the last selected passage may have had multiple items. Therefore, three methods for selecting items with passages were proposed in this study (for both T and P methods): (1) M1: select all items in the passage until the fixed test length is reached; (2) M2: select all items in the selected passages except the last selected passage—the items in the passage are partially or all selected until the fixed test length is reached; (3) M3: select partial items in the passage and stop the selection if the fixed test length is reached. So for method M1, the test length might have been longer than the fixed test length, but methods M2 and M3 had the fixed test length.

The item selection procedures are briefly described below. For all the procedures, the initial abilities are set to $\theta_l = 0$ for $l =$

$1, \cdots, D$. For $j = 2, \cdots, J$, suppose $j-1$ items have been selected. To select the next jth item, suppose the updated ability is $\vec{\theta}^{j-1}$. For each of the procedures, the steps proposed are repeated. Please note that the procedures are for both Bayesian and non-Bayesian; for Bayesian, add Σ^{-1} to the information.

Steps for the T Methods

1. For each passage M_p in the pool (including single items), compute the information at the ability level $\vec{\theta}^{j-1}$,

$$\mathbf{I}_{M_p}(\vec{\theta}^{j-1}) = \sum_{m \in M_p} \frac{(P_{m1} - \beta_{3m})^2(1 - P_{m1})}{P_{m1}(1 - \beta_{mj})^2} \vec{\beta}_{2m} \otimes \vec{\beta}_{2m}$$

   where $m \in M_p$ is for all the items in the testlet or passage M_p.
2. Select the passage M_p such that $\mathbf{I}_{M_p}[0][0]$ has a maximum value (among all the passages in the pool); it is the element in the top right of the information matrix and it measures the precision of the primary ability.
3. Rank order the items in the passage M_p by their informations and select the top n_p items where $n_p = m_p$ or $n_p = Math.min(3, m_p/2 + 1)$ for selection method M1 and M3 respectively, where m_p is the number of items in passage M_p. For selection method M2, $n_p = m_p$ for all except the last selected passage. The process stop if the test length reached the maximum for method M2 and M3.
4. Update ability $\vec{\theta}^{j}$ based on the selected $j - 1 + n_p$ items.

Steps for the P Methods

For the P method, unidimensional IRT is used. The above steps were similarly applied; the number of dimensions is 1.

Steps for the U Methods

U model is a special case when the number of dimensions is 1; the information function is a scalar. The steps are below:

1. For each item m in the pool, compute the information the ability level $\vec{\theta}^{j-1}$,

$$\mathbf{I}_j^m(\vec{\theta}^{j-1}) = \frac{(P_{m1} - \beta_{3m})^2(1 - P_{m1})}{P_{m1}(1 - \beta_{mj})^2} \vec{\beta}_{2m} \otimes \vec{\beta}_{2m}$$

2. Select item $j = m$ such that $\mathbf{I}_j^m(\vec{\theta}^{j-1})$ has a maximum value (among all the items in the pool).
3. Update ability $\vec{\theta}^{j}$ based on the selected j items.

For ability update and final estimates, there are five basic estimation methods that are used in IRT: (a) maximum likelihood estimation (MLE; Lord, 1980, 1984), (b) Maximum a posterior (MAP; Samejima, 1980), (c) Owen's sequential Bayesian (OSB; Owen, 1975), (d) expected a-posteriori (EAP; Bock, 1981), (e) marginal maximum likelihood (MML; Bock, 1981), and (f) weighted maximum likelihood (Warm, 1989). Owen method has been widely used in the process of updating ability after each item selection in unidimensional CAT; the final ability estimate is still non Owen method such as MAP or MLE; the item order affects Owen estimates. Owen method in updating ability is fast and accurate, as it uses only the response for the current selected item and previous ability update. The extension of Owen method in updating the vector ability to the MIRT model would be so much faster than any other methods. However, no research has been conducted.

Studies have been conducted comparing the performance of MAP, weighted MLE, and MLE in unidimensional IRT (Wang and Wang, 2001; Penfield and Bergeron, 2005; Sun et al., 2012). The comparison was extended to the MIRT model in Yao (2018); it was for a fixed length paper and pencil test. It was found that Bayesian method MAP had larger BIAS but smaller SE (standard error) and RMSE. WMLE and MLE had similar results with WMLE performing slightly better than MLE; both had smaller BIAS but larger RMSE and SE compared to MAP. WMLE used the largest computer time in estimating ability and MAP used the smallest computing time. For the simulated data in this research, WMLE for ability update and estimates would require much longer computer time. Therefore, MAP estimation method was used.

Evaluation Criteria

Reliability, BIAS, and SE (standard error) for the abilities were computed average over replications. They were defined as follows: let f_{true} be the value of a function obtained from true parameter and f_l be the value of a function obtained from the estimated parameter from sample l. Here the function f can represent ability parameters. RMSE was calculated by $RMSE = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(f_l - f_{true})^2}$, where n is the number of replications. BIAS was defined by $BIAS = \mid f_{true} - \bar{f} \mid$, where $\bar{f} = \frac{1}{n}\sum_{l=1}^{n} f_l$ was the final estimate. Standard error SE was calculated by $SE = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(f_l - \bar{f})^2}$. Standard error of measurement for content domain l was calculated by $SEM_l = \sqrt{\vec{\alpha}_l \mathbf{I}^{-1} \vec{\alpha}_l^T}$, where $\vec{\alpha}_l = (0, \cdots, 1, 0, \cdots)$ has 1 in the lth element. For Bayesian method, replace $\mathbf{I}^{-1}$ with $(\mathbf{I} + \Sigma^{-1})^{-1}$. Here $\mathbf{I}$ is the information for the test at the estimated abilities. The reliability is the square of the correlations between the estimates and the true values.

Item usage and test overlap rate were computed to evaluate the item pool usage. Test overlap rate is the expected number of common items encountered by two randomly selected examinees divided by the expected test length. It can be derived below:

$$Overlap = \frac{\sum_{j=1}^{T} count_j(count_j - 1)}{J(N(N-1))}. \quad (7)$$

If the exposure rate of an item is bigger than 0.2, then it is defined as overexposed. If the exposure rate of an item is smaller than 0.02, then it is defined as underexposed.

RESULTS

Figures 1–3 shows the reliability, BIAS, empirical SE, and SEM for test length of 10, 20, and 30 for all 9 procedures for the general ability (the first dimension) for methods M1-M3, respectively. The U methods had the smallest SE; the standard errors under the IRT model were overly optimistic. This finding is consistent with previous research (Yen, 1984; Sireci et al., 1991; Wainer and Thissen, 1996; Wainer et al., 2007; Murphy et al., 2010). The U

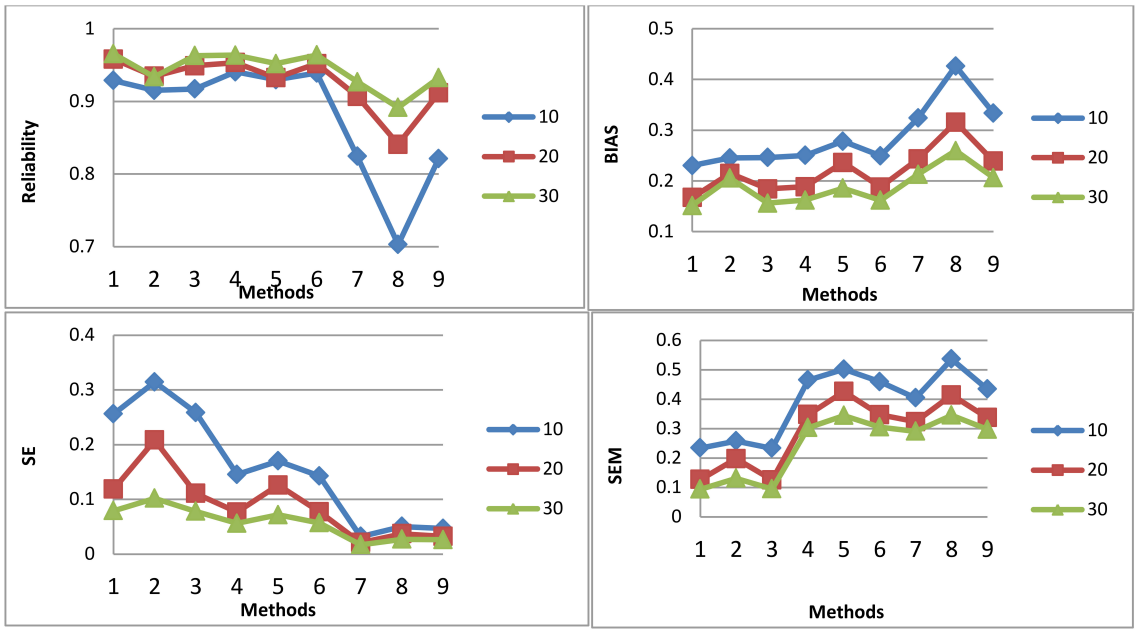

FIGURE 1 | Reliability, BIAS, empirical SE, and SEM for test length of 10, 20, and 30 for all 9 procedures for selection method M1.

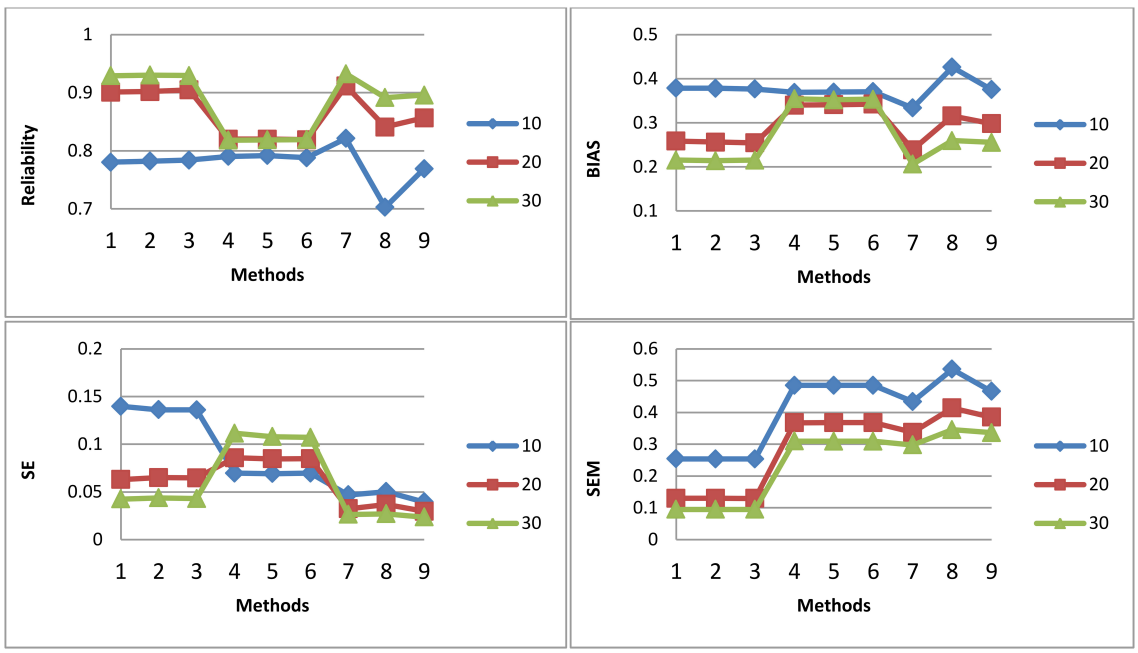

FIGURE 2 | Reliability, BIAS, empirical SE, and SEM for test length of 10, 20, and 30 for all 9 procedures for selection method M2.

models tend to overestimate the precision of the theta estimates when the dependency of items is ignored. For M1, the T model is the best in reliability, BIAS, and SEM; the T and P models may have longer test length than U for M1. However, for M2 and M3, T and P had the same test fixed test length as U. For both M2 and M3, the T model also had the best reliability, BIAS, and SEM. The P model had the smallest reliability and larger BIAS and SEM compared to the T; the dependency of items for the P model was ignored. The P model performed the worst. For the T model, compared to M2, M3 had slightly higher reliability and smaller SE, although the difference was not significant. For M2 and M3, the results for the T and P for the three item exposure

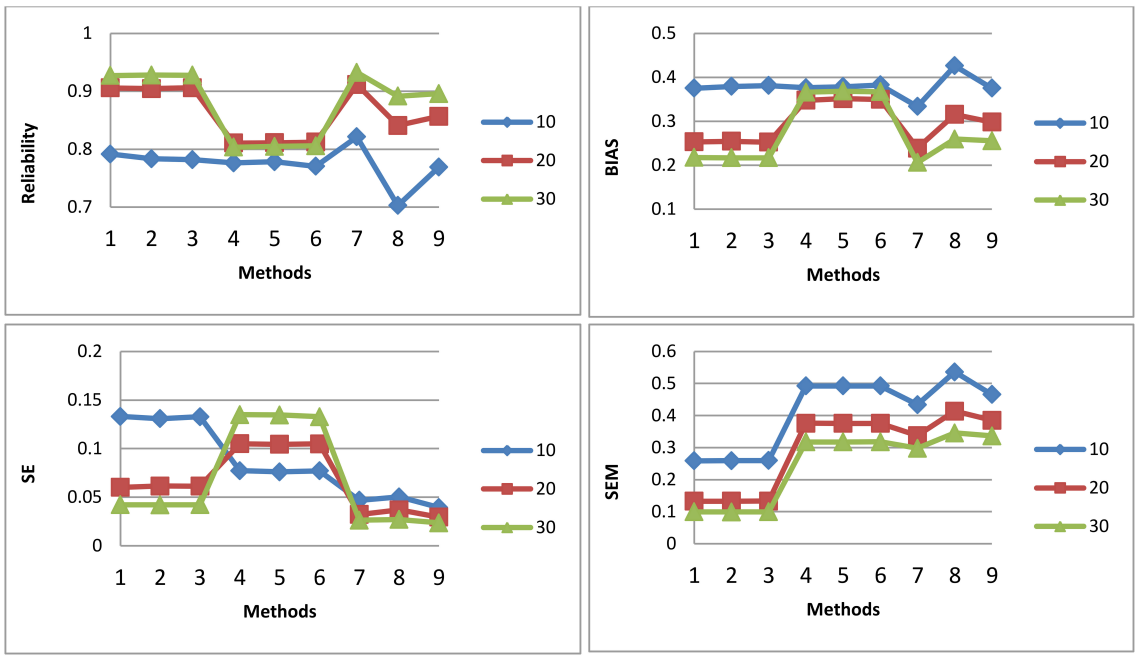

FIGURE 3 | Reliability, BIAS, empirical SE, and SEM for test length of 10, 20, and 30 for all 9 procedures for selection method M3: Note about the x-axis for the 9 procedures: 1, Testlet model; 2, Testlet model with 0.2 as item exposure rate using PI; 3, Testlet model with 1.0 as item exposure rate using PI; 4, Passage model; 5, Passage model with 0.2 as item exposure rate using PI; 6, Passage model with 1.0 as item exposure rate using PI; 7, UIRT model; 8, UIRT model with 0.2 as item exposure rate using PI; 9, UIRT model with 1.0 as item exposure rate using PI.

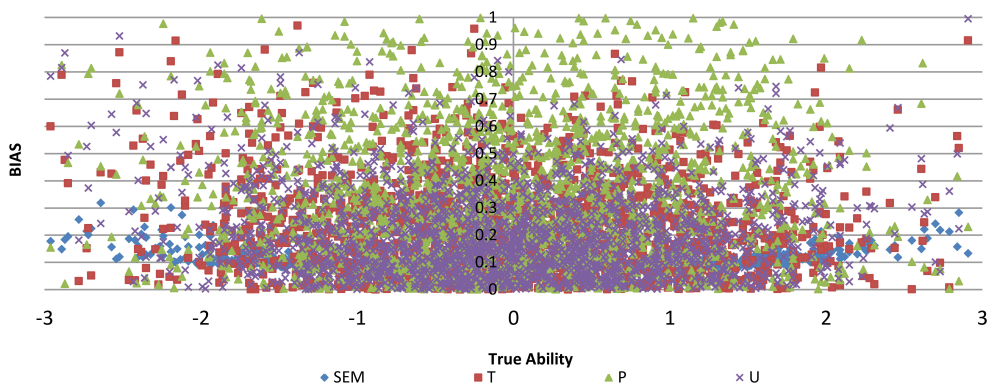

FIGURE 4 | BIAS against true ability for three item selection methods T, P, and U, with no item exposure control for test length of 30 for method M2.

control were similar; this was not surprising, as there were only 38 passages and 16 single items in the pool to be selected. For the U model, the item exposure control had an effect, as there were 100 items to be selected in the pool.

Figure 4 shows the BIAS against true ability for the three item selection methods T, P, and U, with no item exposure control for test length of 30 for M2. It can be seen that T had smaller BIAS for most of the abilities than the P and U models. To see the bias in detail, the 2,000 true abilities were classified into four categories with their means for each category computed; the means for the true ability, BIAS and SEM for each category were computed. **Table 1** displays the results. For all four categories, the T model had the smallest BIAS and SEM. **Table 2** shows the percentage of misclassification rate for the three models T, P, and U. The second column is the category based on the true ability values and the last column is the percentage of misclassification based on the estimated values from the models; for the four cut points in **Table 1**, there were 5 categories. The three models had similar misclassification rates for categories 2, 3, and 4, but there were differences for categories 1 and 5, with T performing the best. Please also note that the T model had the smallest SEM (standard error of measurement).

Figure 5 plots the item usage rate for the 100 items (x-axis) for the three methods T, P, and U for conditions of item exposure rate of 0.2 and 1, and of no exposure control for M2. For item exposure rate of 0.2, all items were being selected and the

TABLE 1 | True ability, BIAS, and SEM for T, P, and U methods for test length of 30 with no item exposure control.

Methods	Category	True Ability	BIAS	SEM
T	1	−1.3088	0.2398	0.1058
P	1	−1.3088	0.3859	0.3173
U	1	−1.3088	0.2351	0.3097
T	2	−0.3321	0.2174	0.0978
P	2	−0.3321	0.3338	0.3182
U	2	−0.3321	0.1908	0.2992
T	3	0.3089	0.2040	0.0926
P	3	0.3089	0.3437	0.3169
U	3	0.3089	0.1837	0.2921
T	4	1.2652	0.2098	0.1019
P	4	1.2652	0.4016	0.3185
U	4	1.2652	0.2154	0.2897

TABLE 2 | Misclassification rate for T, P, and U methods for test length of 30 with no item exposure control.

Methods	Category	Percentage
T	1	2.8
P	1	3.7
U	1	3
T	2	24.55
P	2	24.45
U	2	25.15
T	3	25.25
P	3	25.25
U	3	25.25
T	4	27.55
P	4	27.55
U	4	27.55
T	5	1.1
P	5	2.4
U	5	2.6

maximum exposure rate for items was under 30% for the U model. There were some items that were never selected including testlets with 2 or 3 items within; the maximum exposure rate for items was under 60%, and 85% for the T and P models, respectively. Passage with 4 items within had a higher selected rate, although there were some passages with 3 items within that had a smaller usage rate. For the P model, there were more items, compared to the T and U models, that had 0 or small item usage.

Figure 6 shows the number of selected items, the number of over exposed items, the number of under exposed items, and the item overlap rate for all three test lengths and 9 procedures for M2. Over exposed items were those that had an exposure rate higher of than 0.2; under exposed items were those that had an exposure rate of smaller than 0.02. Overall, the P model had a narrow range of items that were selected. The U model had the smallest overlap rate followed closely by the T model. The T model had a similar item overlap rate as the U model when there was no item exposure control. For test length of 20, the T model and the U model with no item exposure control had similar item usage. Therefore, it is expected that with larger number of testlet items in a pool, the T model should have as good usage of items as the U model.

DISCUSSION

This simulation study compared three models in CAT item selection with passages: (a) the testlet-effect model (T); (b) the passage model(P); and (c) the unidimensional IRT model (U). It was found that TRT (P) and IRT (U) performed similarly in Murphy et al. (2010); however, in this study, the T and P methods were more closer to each other than to the U in reliability and BIAS; the P method here was the same method as the TRT. Both P and U methods tended to overestimate the precision of the theta estimates, as the dependency of items was ignored; both had smaller SE, larger BIAS and SEM. The T model, where the testlet-effect was considered, had the best reliability and the smallest BIAS and SEM. Moreover, the item pool usage for T models could be as good as the U models when there were more testlet items in a pool. Specific or a required number of single items could be selected; this would be helpful when there are not many testlet items. Even though 0 number of required single items were specified in this study, the software can be adopted to handle any input number and is free for the public to use.

Real data was used in deriving the item pools for the models. The UIRT model and the testlet-effect model fit the real data equally well so the comparison between different CAT item selection methods from different models make sense. In real practice, if the UIRT model fits the model the best, then the testlet-effect is not strong, and the UIRT model should be used. Many testing companies still use the U model—for example, the CAT ASVAB military test for content PC(Paragraph Comprehension). The U model is studied here as the baseline. If the testlet-effect model fits the data the best, then it is expected that the testlet-effect model for CAT(T) would be much better than the UIRT model (U) or the passage (P) model. Future study varying the degree of testlet-effect and comparing the performance of T and U and other methods should be conducted; for example, treating items in a testlet as a polytomously

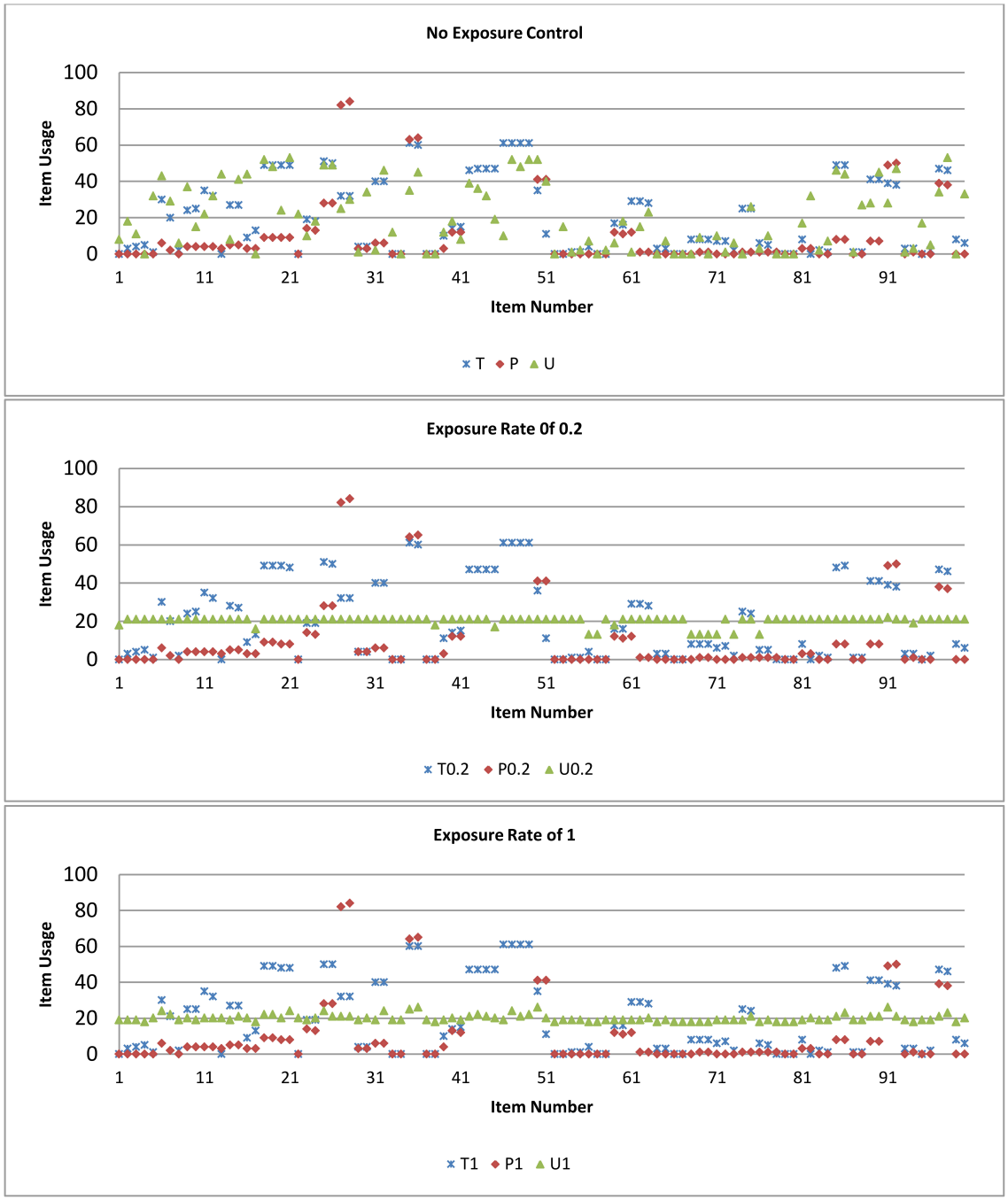

FIGURE 5 | Item usage rate for the three methods T, P, and U for conditions of item exposure rate of 0.2, 1, and no exposure control for test length of 30 for selection method M2.

scored items modeled by the multidimensional generalized two-parameter partial credit model (Yao and Schwarz, 2006).

For this study, the number of passages in the pool was small, therefore, different item exposure controls for both T and P yielded similar results. For pools with larger numbers of passages, the relative performances of T, P, and U would be similar to this study in the case where there is no item exposure control. Overall, the T model is recommended.

Fixed length test was used in this study for the U model and also for the T and P models for M2 and M3. For the passage selection method, partial items or all items in the pool could be selected. Three methods (M1, M2, and M3) were proposed for the T model; other selection criteria can be studied.

This research used only maximum information method (MFI) in selecting items. Other methods such as minimum expected posterior variance (van der Linden and Pashley, 2000),

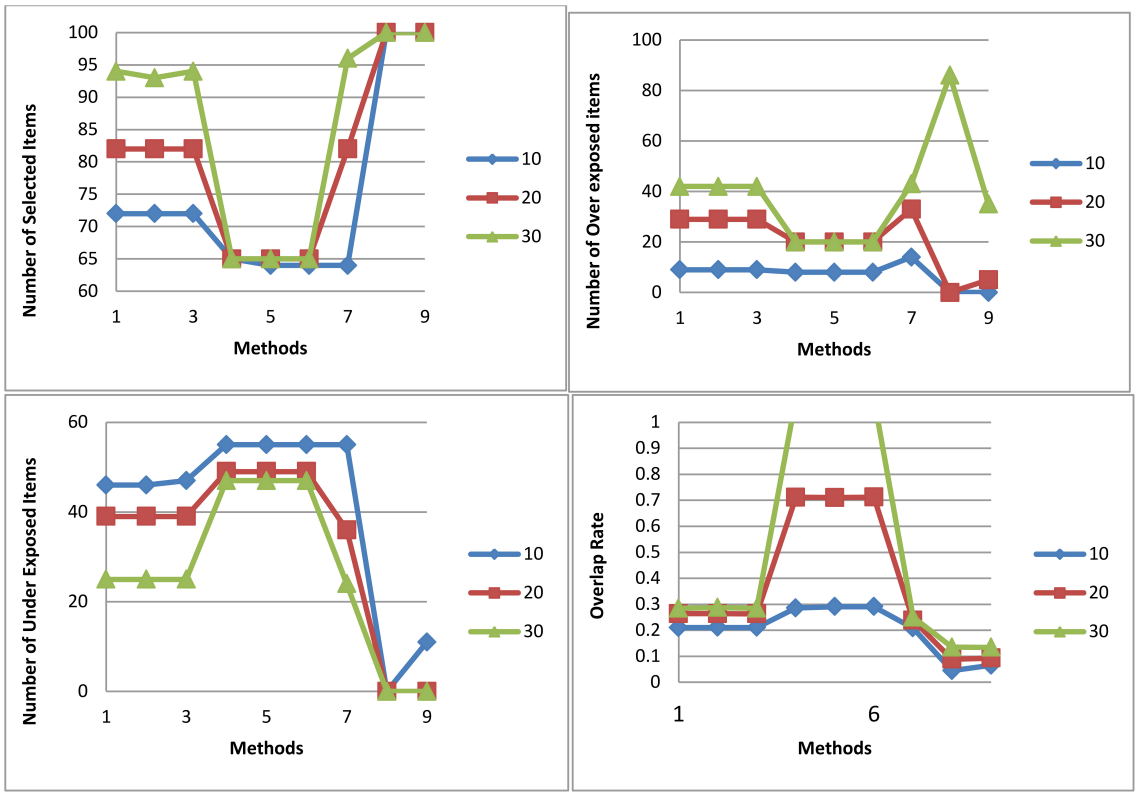

FIGURE 6 | Number of selected items, number of over exposed items, number of under exposed items, and Item overlap rate for all three test lengths and 9 procedures for selection method M2.

Kullback-leibler information (Chang and Ying, 1996), and Volume (Segall, 1996) methods could be studied and compared for T, P, and U models. Similar observations should be expected; however, they might require longer computer time. For CAT using the testlet-effect model, the number of dimensions is the number of testlet plus 1. Therefore, if the number of testlet or cluster in the item pool is larger, then the number of dimensions is larger; other item selection methods besides MFI would be impossible. If the number of testlet or cluster in the item pool is small, then the number of dimensions is small. Most of the items in the pool would be single items. Thus, the relative performance of item selection criteria would be similar to the results as in Yao (2012, 2013).

Other models that consider testlet-effect and content domain information simultaneously can be proposed and studied. Suppose there are K testlets and $D+1$ content domains, with the first dimension as the general dimension measuring general ability, and the rest of the D dimensions are content specific dimensions. Then the model can be $D+1$ dimensional IRT model, and the discrimination parameters are

$$\vec{\beta}_{2j} = (\beta_{2j1}, \beta_{2j1}\gamma_1, \beta_{2j1}\gamma_2, \cdots, \beta_{2j1}\gamma_K) \tag{8}$$

where $\gamma_1, \cdots, \gamma_K$ are testlet-effect parameters for the K testlet. Within each testlet, the ratio of the item general discrimination (β_{2j1}) and the item testlet-effect discriminations is a constant, namely testlet-effect parameter γ_k, where $k \in \{1, \cdots, K\}$. The other item parameters (item difficulty/guessing or alphas) remain the same as the general MIRT model. The number of dimensions and the number of testlets can be different. When the number of dimensions is the number of testlets+1, then it is the regular testlet model that was studied in this paper.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters:application of an EM algorithm. *Psychometrika* 46, 443–459.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlet. *Psychometrika* 64, 153–168.

Chang, H.-H., and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Appl. Psychol. Measur.* 20, 213–229.

Cheng, Y., and Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 62, 369–383. doi: 10.1348/000711008X304376

DeMars, C. (2006). Application of the bi-factor multidimensional item response theory model to Testlet-based tests. *J. Educ. Measur.* 43, 145-168. doi: 10.1111/j.1745-3984.2006.00010.x

Li, Y., Bolt, D. M., and Fu, J. (2006). LA comparison of alternative models for testlets. *Appl. Psychol. Meas.*30, 3–21. doi: 10.1177/0146621605275414

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1984). *Maximum likelihood and Bayesian Parameter Estimation in Item Response Theory* (Research Rep. No. RR-84-30-ONR). Princeton, NJ: Educational testing Service.

Murphy, D. L., Dodd, B. N., and Vaughn, B. K. (2010). A Comparison of item selection techniques for testlets. *Appl. Psychol. Measur.* 34, 424–437. doi: 10.1177/0146621609349804

Owen, R, J. (1975). A Bayesian sequential procedure for quantial response in the context of adaptive mental testing. *J. Am. Stat. Assoc.* 70, 351–356.

Penfield, D. R., and Bergeron, M. J. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Appl. Psychol. Measur.* 29, 218–233. doi: 10.1177/0146621604270412

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Appl. Psychol. Measur.* 21, 25–36.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY:Springer.

Samejima, F. (1980). *Is Bayesian Estimation Proper for Estimating the Individual's Ability* (Research Rep. 80-3). Knoxville, TN: University of Tennessee, Department of Psychology.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika* 61, 331–354.

Sireci, S. G., Wainer, H., and Thissen, D. (1991). On the reliability of testlet-based tests. *J. Educ. Measur.* 28, 237–247.

Sun, S., Tao, J., Chang, H., and Shi, N. (2012). Weighted maximum-a-posteriori estimation in tests composed of dichotomous and polytomous items. *Appl. Psychol. Measur.* 369, 271–290. doi: 10.1177/0146621612446937

van der Linden, W., and Pashley, P. J. (2000). "Item selection and ability estimation," in *Computerized Adaptive Testing: The031 and Practice*, eds W. J. van der Linden and C. A. W. Glas (Nonvell, MA: Kluwer Academic Publishers), 1–25.

van der Linden, W. J., and Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *J. Educ. Behav. Stat.* 29, 273–291.doi: 10.3102/10769986029003273

van der Linden, W. J., and Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *J. Educ. Behav. Stat.* 32, 398–417. doi: 10.3102/1076998606298044

Veldkamp, B. P., and van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika* 67, 575–588. doi: 10.1007/BF02295132

Wainer, H., Bradlow, E. T., and Wang, X. (2007). Testlet response theory and its applications. *Psychometrika* 74. 555–557. doi: 10.1007/s11336-009-9120-5

Wainer, H., and Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educ. Measur. Issues Prac.* 15, 22–29.

Wang, S., and Wang, T. (2001). Precision of Warm's weighted likelihood estimation of ability for a polyto-mous models in computer adaptive testing. *Appl. Psychol. Measur.* 25, 317–331. doi: 10.1177/014662101220 32163

Wang, W.-C., and Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Appl. Psychol. Measur.* 29, 296–318. doi: 10.1177/0146621605276281

Waniner, H., Bradlow, E. T., and Du, Z. (2000). "Testlet response theory, an analog for the 3PL model useful in testlet-based adaptive testing," in *Computerized Adaptive Testing: Theory and Practice*, eds W. J. van der Linden & C. A. W. Glas (Dordrecht: Kluwer), 245–269.

Warm, A. T.(1989). Weighted Likelihood estimation of ability in Item response theory. *Psyvhomrtrika* 54, 427–450.

Yao, L. (2003). *BMIRT: Bayesian Multivariate Item Response Theory [Computer software]*. Monterey, CA: DMDC.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: theory and applications. *Psychometrika* 77, 495–523. doi: 10.1007/s11336-012-9265-5

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Appl. Psychol. Measur.* 37, 3–23. doi: 10.1177/0146621612455687

Yao, L. (2018). Comparing methods for estimating the abilities for the multidimensional models of mixed item types. *Commun. Stat.* 47, 74–91. doi: 10.1080/03610918.2016.1277749

Yao, L., and Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Appl. Psychol. Measur.* 30, 469–492. doi: 10.1177/0146621605 284537

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Measur.* 8, 125–145.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Measur.* 30, 187–213.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.