Article

Location-Matching Adaptive Testing for Polytomous Technology-Enhanced Items

Applied Psychological Measurement
2024, Vol. 48(1-2) 57–76
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216241227548
journals.sagepub.com/home/apm
Sage

Hyeon-Ah Kang[1], Gregory Arbet[1], Joe Betts[2], and William Muntean[2]

Abstract

The article presents adaptive testing strategies for polytomously scored technology-enhanced innovative items. We investigate item selection methods that match examinee's ability levels in location and explore ways to leverage test-taking speeds during item selection. Existing approaches to selecting polytomous items are mostly based on information measures and tend to experience an item pool usage problem. In this study, we introduce location indices for polytomous items and show that location-matched item selection significantly improves the usage problem and achieves more diverse item sampling. We also contemplate matching items' time intensities so that testing times can be regulated across the examinees. Numerical experiment from Monte Carlo simulation suggests that location-matched item selection achieves significantly better and more balanced item pool usage. Leveraging working speed in item selection distinctly reduced the average testing times as well as variation across the examinees. Both the procedures incurred marginal measurement cost (e.g., precision and efficiency) and yet showed significant improvement in the administrative outcomes. The experiment in two test settings also suggested that the procedures can lead to different administrative gains depending on the test design.

Keywords

technology-enhanced items, polytomous scoring, response time, computerized adaptive testing, variable-length computer-adaptive classification testing

Introduction

Many testing programs nowadays use computers for test delivery and seek to include technology-enhanced items (TEIs). TEIs are the items that make use of computer functions and take nontraditional innovative formats. Items may use digital tools to present information

[1] University of Texas at Austin, TX, USA
[2] National Council of State Boards of Nursing, IL, USA

Corresponding Author:
Hyeon-Ah Kang, University of Texas at Austin, 1912 Speedway, Stop D5800, Austin, TX 78712, USA.
Email: hkang@austin.utexas.edu

efficiently (e.g., audio, graphics, video, and simulation) or ask dynamic user interaction to assess processing skills or complex knowledge (e.g., drag-and-drop, hot spot, cloze, and sequencing). Afforded by modern technology, TEIs demonstrate greater fidelity to real-world problems and can yield better measurement outcomes than traditional text-based items (e.g., enhanced construct validity and content coverage, reduced guessing).

While TEIs hold great promise for future assessments, much work is needed for reaping the full benefits. One area of special importance is item administration. In applied settings, TEIs are commonly presented with the existing items in small numbers. Items can be interspersed manually at desired locations or seeded automatically on a statistical criterion. In the latter case of stochastic assignment, common practice is to use an information measure. Computerized adaptive testing (CAT), for example, evaluates information that candidate items provide for the examinee's latent proficiency and administers an item that provides maximal information at the currently estimated ability level. The maximum information (MI) criterion offers an optimal design for maximizing the measurement efficiency (i.e., minimizes the number of items needed for achieving the same measurement precision) and has been widely adopted in many operational testing programs.

The MI criterion, despite the virtue of measurement efficiency, can however entail an item usage problem. The criterion tends to select informative items exceedingly frequently while seldom or never choosing the items that contain relatively less information. Applying the MI criterion to TEIs faces the same usage problem but it occurs at a greater cost. Developing TEIs is more costly than developing standard text-based items, and underusing items will make the extended efforts fruitless while depriving the chance for future refinement. On the other hand, overuse of items not only exposes the items to security breach but it can also lead to inordinately time-exacting tests as informative items tend to require longer times.

Recognizing the concern on the usage problem of the information-based testing, the current study seeks to investigate item selection methods that can be alternatively used for TEIs. In real settings, TEIs are typically administered with regular items, and greater diversity in item sampling may be more preferred even when the items do not provide maximum information. For example, TEIs that tap various content areas, demonstrate high authenticity, or provide positive user experience may have operational values even when they provide less-than-optimal information. With this being noted, the current study explores item selection methods for TEIs that can achieve greater sampling diversity on a statistical criterion.[1] Our approach to item selection is to match items' location to examinees' ability levels such that items can be sampled in a wide variety, resembling the ability distribution. Since examinees typically have a wide range of ability levels, we surmise that location-matched item selection would achieve greater diversity in item sampling and lead to more balanced item pool usage.

The idea of location-matching item selection is not new and has been widely practiced in binary-response adaptive testing through the *b*-matching criterion. The *b*-matched selection is less greedy than the information-based selection and tends to make more exhaustive use of item pools. One issue with applying this criterion to TEIs is that TEIs are typically scored polytomously (Betts, Muntean, Kim, & Kao, 2021; Jiao et al., 2012; Kang, Han, Betts, & Muntean, 2022), and there is no specific index that can characterize the location of an item. In this study, we introduce location indices that describe the location of a polytomous item on the ability continuum and investigate probable testing outcomes of the location-matched item selection through Monte Carlo simulation. While formulating the selection criterion, we also contemplate utilizing time information so that the selection decision can be informed of examinees' time-wise behavior and can accommodate working speeds.

We note that the current literature contains a wealth of studies that examine the item selection methods for polytomous-response CAT (e.g., Pastor et al., 2002; van Rijn et al., 2002; Veldkamp, 2003). As alluded to earlier, these studies mostly applied the information measures (e.g., Fisher information and Kullback–Leibler divergence) that are susceptible to the usage problem. The current study is aimed to explore other possible approaches to selecting polytomous items, giving special consideration to the administration of TEIs.

In the sections that follow, we expound on the proposed idea, details, and performance of the suggested item selection methods. In Section 2, we discuss measurement models for item responses and response times. We then present an inferential framework that jointly estimates parameters of the measurement models. Section 3 discusses adaptive item selection strategies suggested for polytomous TEIs. It begins with the MI criterion and a variant that integrates response time information. The section continues with the location-matching (LM) selection approach, followed by speed-moderated LM criteria. Sections 4 and 5 present simulation studies under two test settings: computerized adaptive testing (CAT) and variable-length computer-adaptive classification testing (VL-CCT). The paper concludes in Section 6 with a summary of findings, implications, and future research directions.

Model and Inference

Measurement Models

The study applies the generalized partial credit model (GPCM; Muraki, 1992) to model item response scores and the log-normal model (LNM; van der Linden, 2006) to describe response time behavior. GPCM describes the probability of an item score as

$$P_{jk} = \Pr\left(X_j = k \middle| \theta\right) = \frac{\exp\left(\sum_{l=0}^{k} a_j\left(\theta - b_{jl}\right)\right)}{\sum_{h=0}^{K_j} \exp\left(\sum_{l=0}^{h} a_j\left(\theta - b_{jl}\right)\right)},$$

where X_j ($\in \{0, 1, \ldots, K_j\}$) denotes an examinee's response score on item j, θ models examinee's latent ability, and a_j and b_{jk} denote item parameters each representing discriminating power and difficulties of the response categories (k = 1, … , K_j; $b_{j0} \equiv 0$).

Response time on an item is modeled as

$$f\left(T_j = t_j \middle| \tau; \alpha_j, \beta_j\right) = \frac{\alpha_j}{t_j\sqrt{2\pi}} \exp\left(-\frac{\alpha_j^2}{2}\left(\log t_j - \left(\beta_j - \tau\right)\right)^2\right),$$

where T_j ($\in\mathbb{R}^+$) denotes a time variable, τ models examinee's latent working speed, and α_j and β_j are item parameters that show time-discriminating power and time intensity of the item. We note that GPCM and LNM are chosen as baseline measurement models for the study; extensions to other models are straightforward.

Joint Item Calibration

The parameters of GPCM and LNM can be estimated jointly under the hierarchical framework (van der Linden, 2007). This study extends the marginal maximum likelihood estimation of Kang, Zheng, and Chang (2021) to draw joint inference from GPCM and LNM. Let ξ contain structural parameters of the joint model (i.e., item parameters and person hyperparameters). The marginal likelihood of ξ is evaluated as

$$\mathcal{L}(\xi|\mathbf{Y}) = \prod_{i=1}^{N} p(\boldsymbol{y}_i|\xi) = \prod_{i=1}^{N} \int p(\boldsymbol{y}_i|\boldsymbol{\eta},\xi) \quad p(\boldsymbol{\eta}|\xi)\, d\boldsymbol{\eta},$$

where $\mathbf{Y} = (\boldsymbol{y}_i: i = 1, \ldots, N)$, $\boldsymbol{y}_i = ((\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij})^\top : j = 1, \ldots, J)$, and $\boldsymbol{\eta} = (\theta, \tau)$. N and J each denote the number of examinees and items in the calibration sample. Given the marginal likelihood, the structural parameters are estimated as

$$\hat{\xi} : = \arg\max_{\xi} \log \mathcal{L}(\xi \mid \mathbf{Y}), \tag{1}$$

giving a marginal maximum likelihood estimator.

In the case that priors are available for the item parameters, ξ can be estimated via marginal maximum a posteriori. Let us apply a normal prior on the transformed item parameters, (log a, b, log α, β), and ξ^* contain the corresponding structural parameters. The estimator can be then obtained as

$$\hat{\xi} : = \arg\max_{\xi} \log p(\xi^* \mid \mathbf{Y}) \propto \arg\max_{\xi} \log p(\xi^* \mid \mathbf{Y}) + \log p(\xi^* \mid \Xi), \tag{2}$$

where Ξ contains hyperparameters of the item parameters. Appendix A provides details of the estimating functions. Software programs that implement the estimation can be accessed via an open source platform.

Trait Inference Under the Joint Framework

Once item parameters are estimated with sufficient precision, examinees' trait levels can be estimated based on the known item parameters. This study applies expected a posteriori (EAP) for the first item and maximum a posteriori (MAP) afterward. Applying Bayes prior p (θ, τ), a posterior probability of trait values is evaluated as

$$p(\theta, \tau \mid \boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{\Omega}) = \frac{\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{t}_i \mid \theta, \tau)\, p(\theta, \tau \mid \boldsymbol{\Omega})}{p(\boldsymbol{x}_i, \boldsymbol{t}_i)}, \tag{3}$$

where $\boldsymbol{\Omega}$ contains hyperparameters of the trait parameters and $\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{t}_i \mid \theta, \tau) = \prod_{j=1}^{J} p(x_{ij} \mid \theta) p(t_{ij} \mid \tau)$ under the conditional independence (van der Linden & Glas, 2010). EAP and MAP are then obtained as the mean and mode of (3).

Item Selection Methods

The study investigates two criteria for selecting polytomous TEIs: (i) location matching and (ii) speed leverage. The location-matching selects items that match the examinee's ability level in location. Since polytomous items have no specific index that represents location, we suggest two indices from Ali et al. (2015) that have potential for practical use. Both the indices characterize relative location of an item on the ability continuum. The item selection based on these indices can be seen as adapting item's overall difficulty to an examinee's ability level. As explained earlier, since examinees typically have a wide variety of ability levels, we anticipate that the location-matched item selection achieves greater diversity in item sampling than the maximum information criterion.

The attempt to leverage test-taking speed in item selection is guided by our observation from real assessment data. TEIs in real assessments vary substantially depending on the presentation format, answer type, and location on the test, and they show markedly different response time patterns. For example, items asking typed answers tend to require more time than the items asking mouse-clicked or dragged answers. TEIs that appear later tend to show shorter processing times as examinees become familiar with the item formats. These items may show little difference in the item response parameters (and thus in the item information) but can differ substantially in the time-wise parameters. Item selection that takes into account these distinct characteristics can not only help regulate examinees' testing times but can also improve item pool usage as it can alleviate reliance on the item information.

With the above new approaches in view, below we present new selection criteria that match items' location in overall difficulty and time intensity. We begin with existing methods that provide reference conditions—the MI criterion and its variant that leverages response time—and then continue with the proposed LM criteria and variations that leverage working speed. For illustration, this study uses Fisher information as a representative example of the information-based item selection. We leave extensions to, and comparison with, other information measures (e.g., Kullback–Leibler divergence, Shannon entropy, and mutual information) for future research.

Information-Based Item Selection

Maximum Information (MI). The MI criterion selects an item that provides maximum information about examinee's latent ability. Suppose at stage m (i.e., the mth item on the test) an item is selected from a pool of eligible items. The MI criterion evaluates item information based on the examinee's provisional ability estimate, $\widehat{\theta}$, and administers the item that provides maximum information

$$j_{m+1} := \underset{j \in \mathcal{R}_m}{\arg\max}\, I_j(\widehat{\theta}), \tag{4}$$

where j_{m+1} is the $(m+1)$th item to be selected, $\mathcal{R}_m$ denotes the set of eligible items remained after administering the m items, and $I_j(\widehat{\theta})$ evaluates information that the prospective item j provides at $\widehat{\theta}$. In the case of GPCM, the item information is calculated as

$$I_j(\theta) = a_j^2 \left(\sum_{k=0}^{K_j} k^2 P_{jk}(\theta) - \left(\sum_{k=0}^{K_j} k P_{jk}(\theta) \right)^2 \right),$$

where $P_{jk}(\cdot)$ is the item category response function for item j category k.

Maximum Information Per Time Unit (MIT). The MI criterion provides an optimal item selection design in the sense that it can achieve the same measurement precision using a minimal number of items (i.e., length-wise measurement efficiency). It however gives little consideration to the time needed for answering an item. In real settings, items that contain large information often demand substantial time effort (van der Linden et al., 1999), and sole focus on the item information can lead to time-exacting tests. The information-centered item selection in particular can handicap capable examinees as the items that are matched to the high ability levels tend to be time-intensive (van der Linden & van Krimpen-Stoop, 2003).

With the notion that item information correlates with the time intensity, several studies considered modulating response times in item selection (e.g., Cheng et al., 2017; Choe et al., 2018; Fan et al., 2012). A pioneering work in this attempt is Fan et al. (2012), which proposed to select an item that maximizes information per time unit (MIT). The selection criterion uses the standard information measure but inversely weighs the expected response time of the prospective item.

Again suppose that an item is sampled at stage $m + 1$. The MIT criterion then selects the next item as

$$j_{m+1} : = \underset{j \in \mathcal{R}_m}{\arg\max} \frac{I_j(\widehat{\theta})}{(E[T_i \quad | \quad \widehat{\tau}])^{\omega}}, \tag{5}$$

where $E[T_i \,|\, \widehat{\tau}] = \exp\left(\beta_j - \widehat{\tau} + \frac{1}{2\alpha_j^2}\right)$ evaluates the time expected to take for answering the item j when examinee is estimated to work at speed estimate $\widehat{\tau}$. Observe that (5) allows different weights in time leverage, ω ($\in$ [0, 1]). While the original criterion assumed full time weight, this study applies a differential weighting scheme to modulate the impact of the time leverage and investigates effects of the different time weights on the item selection and test outcomes.

Location-Matched Item Selection

Both the MI and MIT rely on the item information measure and are subject to the item pool usage problem. In this study, we explore a location-matching approach that could lead to more balanced item pool usage. Two indices are considered to determine the location of a polytomous item: (i) an average of step difficulties and (ii) an intermediate ability point that leads to a half item score (i.e., $K_j/2$).[2] The average point models the overall difficulty of an item through the average of the step difficulty parameters of the response categories

$$\xi_j = \frac{1}{K_j} \sum_{k=1}^{K_j} b_{jk}, \tag{6}$$

where b_{jk} is the difficulty of the response category k, and K_j is the maximum item score. The intermediate point is the point at which examinee is expected to score a half point of the item score

$$\xi_j = \theta\,(\in\Theta) : E\left[X_j\right] = \sum_{k=0}^{K_j} kP_{jk}(\theta) = \frac{K_j}{2}. \tag{7}$$

The root of (7) can be found via Newton–Raphson iteration as

$$\xi_j = \theta : f(\theta) = \sum_{k=0}^{K_j} kP_{jk}(\theta) - \frac{K_j}{2} = 0$$

with

$$\theta_{n+1} = \theta_n - \frac{f(\theta_n)}{f'(\theta_n)} = \theta_n - \frac{\left(\sum_{k=0}^{K_j} kP_{jk}\right) - \frac{K_j}{2}}{a_j \sum_{k=0}^{K_j} kP_{jk}\left(k - \sum_{h=0}^{K_j} hP_{jh}\right)}$$

giving an updated approximate. A selection criterion based on the location index is then formulated as a minimization problem that searches an item with minimum distance from the exaiminee's provisional ability estimate

$$j_{m+1} : = \underset{j \in \mathcal{R}_m}{\arg\min} \left|\xi_j - \widehat{\theta}\right|. \tag{8}$$

Setting ξ_j as (8) leads to an averaging matching (AM) criterion; setting at (7) leads to an intermediate matching (IM) criterion. The AM selection can be seen as adapting items' overall difficulties to examinee's ability. The rationale for the IM selection originates from the classical dichotomous CAT that defines the item location as the point that leads to Pr $(X = 1\,|\theta) = .5$ (Lord, 1970). Repeated sampling toward (7) leads to a reasonable guess for the examinee's true ability level (Chang, 2015).

As was done in the MIT, the LM selection can integrate information from the response times. Instead of weighing the expected time, however, the criterion leverages examinee's working speed to be in line with the location matching:

$$j_{m+1} : = \underset{j \in \mathcal{R}_m}{\arg\min} \left|\xi_j - \widehat{\theta}\right| \cdot \left|\beta_j - \widehat{\tau}\right|^{\omega}. \tag{9}$$

The criterion (9) selects an item that minimizes distance in both the ability and speed continua.[3] Since items are selected from wide ranges of item-difficulty and time-intensity parameters that match the examinee's ability and working speed, we anticipate that the criterion would achieve more diverse sampling than MI(T) that favors informative or facile items. Again note that (9) permits different weights in the speed leverage and the effect of the added leverage can be modulated during the item selection.

Simulation Study I: CAT

Monte Carlo simulation studies were conducted to evaluate the performance of the selection methods. The simulation was performed in two test settings: (i) fixed-length computerized adaptive testing (CAT) and (ii) variable-length computer-adaptive classification testing (VL-CCT). Many operational testing programs adopt these testing modes as a main delivery mode, and the choice of an item selection method can carry different implications. In this section, we present the CAT study, giving close attention to the trait recovery, item pool usage, and testing time. The following section presents the study under VL-CCT, examining final test lengths, classification performance, item pool usage, and testing times.

Design

Model. The study used the two models in Section 2.1 as main measurement models—GPCM for item response scores and LNM for response times.

Item Pool. For creating item pools, we simulated pseudo items assuming moderate correlation between the parameter domains, ρ = .3 (Cheng et al., 2017; Klein Entink et al., 2009).[4] We sampled (log a_j, b_{jk}, log α_j, β_j) from a multivariate normal distribution with means $(-.043, 0, -.043, 0)$ and variances $(.086, 1, .086, 1)$ such that the item parameters have means of $(1, 0, 1, 0)$ and variance of (.09, 1, .09, 1) on the original metric. Each item pool contained a set of 200 polytomous items that differ in the maximum item scores $(2 \leq K_j \leq 6)$.[5]

It may be relevant to note that some previous studies applied different distributions for generating step difficulty parameters (e.g., Penfield, 2006; Sun et al., 2012). While this approach helps separate response score categories, it makes it difficult to simulate correlation between the parameter domains. Correlation between the item parameters is one of the important aspects that needs to be considered in the current simulation study because operational items in real assessments often exhibit nonzero correlation (van der Linden et al., 2010) and it affects item selection as well as item pool usage. In addition, since our simulation study assumed different ranges of item scores, it is more efficient to simulate items using the same generating distribution and exclude the items that have calibration issues. This was the strategy applied in this study. We simulated items in large numbers, calibrated prior to CAT simulation, and randomly selected 200 items among those that converged successfully. This strategy better mimics the real test practice and the simulation results will give closer approximations to real outcomes as they embody both the calibration and measurement error. The final item pools used in the CAT simulation contained constant numbers of 200 items that converged successfully in precalibration. For calibrating the items, we applied spiraled linking design with $N = 1000$ calibration samples per item. The item parameters were estimated jointly via marginal maximum a posteriori estimator (see Model and Inference).

CAT. Upon creating item pools, we simulated CAT mimicking real test settings. We drew random samples of $N = 1000$ examinees from a bivariate normal distribution ($\boldsymbol{\mu} = \mathbf{0}, \sigma_\theta^2 = \sigma_\tau^2 = 1, \rho_{\theta\tau} = .3$) and administered 30 items adaptively applying the MI, AM, and IM criteria. When the item selection utilized the information from response times, the time-moderated term was weighted differentially (ω = .25, .50, .75, 1.00) so that the impact of the time/speed leverage can be modulated at different degrees. The selection of the first item was made randomly in all conditions within an exposure constraint.

In applied settings, testing programs typically arrange exposure control to regulate item overuse (e.g., Chang & Ying, 1999; Kingsbury & Zara, 1989; Sympson & Hetter, 1985) or to balance content coverage (e.g., Cheng & Chang, 2009; van der Linden, 2000). Deploying such procedures in the current setting will shield the original usage problem and limit the conclusions to the chosen method. To make our discussion adequately relevant to the current issue and to broader applications, we apply a rather simple exposure control scheme that constrains maximum item exposure rate. We define exposure rate as the frequency of item use over the number of test takers and constrain the maximum rate at .20 such that items can be used up to 20% of the sample. This approach can demonstrate the gravity of the usage problem and will provide a rough and yet realistic projection for real settings. For those situations with more refined or relaxed conditions (e.g., randomesque and higher maximum exposure rate), see Appendix B that provides supplementary simulation.

Trait Estimation. During CAT, examinees' trait levels were estimated by applying standard estimators. When tests were operated based on the responses only, we applied a maximum likelihood or EAP estimator under GPCM.[6] When the information from the response times was utilized, MAP or EAP (first item) was applied under the GPCM and LNM.

Replication. All simulation conditions were replicated 100 times with unique generating parameter sets. Results from the replications were summarized by average and standard deviation (SD) of the evaluation statistics which are discussed below.

Evaluation

Simulation results were evaluated in three aspects: (i) accuracy of final ability estimates, (ii) item pool usage, and (iii) testing time. The trait recovery was evaluated by root mean squared error (RMSE), absolute bias (AbsBias), correlation (Cor) between the true and estimated ability values, and average standard error (SE). The item pool usage was examined by χ^2 (Chang & Ying, 1999), test overlap rate, and percentages of items that reached the maximum exposure (i.e., retired items) and that are underutilized (less than 1%). The results related to the testing times were examined by descriptive statistics: average, maximum, and SD of log testing times.[7] When examining the evaluation statistics, we conducted significance tests and analyzed the variance of the outcome statistics to draw distinction between the selection methods and determine the effect of the design variables. Significance of difference was evaluated at $\alpha < .001$ due to the large number of replications. Significance of an effect was assessed following the convention (Cohen, 1988)—the partial η^2 less than .01 as small effect size, between .06 and .14 as medium, and greater than .14 as large effect size.

Results

Trait Recovery. Table 1 reports results related to the trait recovery. The figures from the ω = 0 condition (i.e., no time leverage) suggest that the selection methods overall performed adequately well. The error rates were kept small (RMSE = .174, AbsBias = .136 on average); correlation with the generating trait values was sufficiently high (.985 on average). Comparison between the selection methods revealed that location-matching criteria led to slightly degenerated measurement precision ($\eta^2 < .055$). The magnitude of degradation was however barely discernible and the differences with MI were mostly less than third decimals. Between the two LM criteria, IM performed marginally better ($p > .048$). Leveraging response times similarly resulted in loss of measurement precision ($\eta^2 > .275$). As we compare the ω = 0 and 1 conditions, we find that the trait values estimated under the full time weight contained larger RMSEs (Δ = .009 on average), were more biased (Δ = .007), and had weaker correlation with the generating parameters ($\Delta = -.001$). Note that both the location matching and time leverage depart from the optimal item selection, and they are expected to yield suboptimal measurement outcomes. The difference in the measurement precision was however generally small and mostly occurred at the third decimal place in the current simulation. Table 1 also shows that the effect of the time leverage can be adjusted with the use of a weighting scheme ($\eta^2 > .187$). Although the trends slightly differed depending on the selection method, zero or a quarter weight generally led to best measurement outcomes.

Item Pool Use. Table 2 presents evaluation statistics relating to the item pool usage. Expectedly, a clear pattern was observed among the selection methods. MI used the item pools most lopsidedly, showing average χ^2 of 7.333 and test overlap rate of 18.58%. It also drove out the large portion of items to the utmost use, exposing 62.08% of items to maximum capacity. The LM criteria, by contrast, showed substantially better and more balanced pool usage ($\eta^2 > .626$). Both the χ^2 and overlap rate decreased significantly to 3.157 and 16.50%. The LM selection also led to more exhaustive pool usage, utilizing almost all items and reducing the percentage of maximally used

Table 1. Trait Recovery in CAT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
RMSE	.173	.172	.172	.175	.177	.174	.175	.178	.182	.185	.174	.173	.177	.181	.184
AbsBias	.134	.134	.134	.137	.139	.136	.137	.140	.142	.144	.136	.136	.139	.141	.144
Cor	.985	.985	.985	.985	.985	.985	.985	.984	.984	.983	.985	.985	.984	.984	.983
SE	.166	.164	.164	.167	.169	.168	.168	.172	.175	.178	.167	.167	.170	.174	.177

Note. RMSE: Root mean squared error. AbsBias: Absolute bias. Cor: Correlation. SE: Standard error. Underlines indicate the best result within each selection method.

Table 2. Item Pool Usage in CAT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
χ^2	7.33	7.36	7.42	7.48	7.59	3.18	1.95	1.40	1.35	1.52	3.14	1.92	1.38	1.34	1.51
Overlap	.186	.186	.186	.187	.187	.165	.159	.156	.156	.157	.165	.159	.156	.156	.157
% Max	.621	.621	.625	.630	.635	.350	.182	.085	.061	.061	.341	.175	.082	.056	.059
% Under	.010	.011	.011	.011	.012	.000	0	0	0	0	0	0	0	0	0

Note. Overlap: Average test overlap rate between two random examinees. % Max: Percentage of items that reached maximum exposure (20% of the sample).% Under: Percentage of under-used items (less than 1%). Underlines indicate the best result within each selection method.

items to 34.56%. Between the two LM criteria, IM achieved slightly better pool utilization ($p > .146$).

The impact of leveraging time/speed differed by the selection method. Weighting response times in MI led to more skewed pool use ($\Delta\,\chi^2 = .255$, Δ Overlap = .13%, Δ Max = 1.39%, Δ Under = .23%). In LM, the speed leverage induced more even pool usage ($\Delta\,\chi^2 = -1.645$, Δ Overlap = −.82%, Δ Max = −28.56%). The different patterns in the selection methods seemed to be related to the characteristics of the items selected by the criteria (see figures in Appendix C for example illustration). In MI, introducing time leverage led to frequent use of less time-intensive items, slanting the item exposure distribution even more. The speed leverage in the LM criteria led to more diverse item sampling across the different location values and βs, entailing a more normal-like exposure distribution. It appears that using timing information generally conduces to more balanced item pool usage in LM, whereas it intensifies the asymmetry in MI.

The trends across the varying time weights similarly differed according to the selection criterion. In MI, increase in the time weight resulted in more imbalanced pool use ($\eta^2 > .158$ except for $\eta^2_{\text{under}} = .009$); in LM, increasing speed weight generally led to more even use ($\eta^2 > .545$ except for $\eta^2_{\text{under}} = .002$) though there seemed to be a midpoint where the selection achieved best pool usage (e.g., $\omega = .75$). The SDs of the evaluation statistics also indicated that the selection criteria differed in stability. Increase in the time weight tended to induce greater variability in MI, whereas it led to smaller variation and more stable performance in LM.

Testing Time. Table 3 reports statistics related to the testing times. The results show that considering response times in item selection indeed reduced the test completion times ($\eta^2 > .128$). The time-leveraged item selection showed significantly shorter average and maximum testing times compared with the regular response-based selection (Δ Avg = −.287, Δ Max = −.455). Among the selection criteria, MI showed the most reduction, shortening the average and maximum testing times by 12.82% and 6.05%, respectively. The LM criteria saved 3.38% (Avg) and 5.36% (Max) on average. The variation in testing times showed somewhat different patterns depending on the selection method. In MI, considering response times led to greater variation across the examinees ($\eta^2 = .683$). Leveraging working speed in LM homogenized the test completion times, reducing the SD by 31.15% on average ($\eta^2 = .972$). The trends across the different time weights were generally consistent with those reported above—the larger the ω, the greater the reduction in testing times ($\eta^2 > .068$). Table 3 also suggests that although LM was not aimed for testing time efficiency, the two selection methods showed shorter average testing times than MI ($\eta^2 = .080$). We believe that this may be related to the reduced emphasis on the item information. Informative items tended to have high discrimination power, great difficulty, and high time intensity, and shifting the

Table 3. Logarithm of Testing Times in CAT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
Avg	4.42	4.26	4.08	3.94	3.85	4.37	4.33	4.29	4.25	4.22	4.36	4.33	4.29	4.25	4.22
Max	8.11	7.88	7.77	7.66	7.62	8.15	8.01	7.96	7.86	7.72	8.20	7.96	7.89	7.78	7.75
SD	1.05	1.06	1.09	1.12	1.14	1.05	.94	.86	.79	.73	1.06	.94	.86	.79	.73

Note. Avg = Average of log total testing time. Max = Maximum of log testing time. SD = Standard deviation of log testing time. Underlines indicate the best result within each selection method.

focus to item location may have led to shorter testing times overall. Between the two LM methods, AM showed marginally shorter testing times (p > .538).

Simulation Study II: VL-CCT

The second simulation study examined the performance of the selection methods in VL-CCT. Since the choice of an item selection criterion affects test length, we give close attention to the mediation effects of the test length.

Design

VL-CCT. The study applied similar settings with the preceding study. Using the precalibrated item pools, samples of $N = 1000$ simulees were tested based on the items that were adaptively selected according to the MI, AM, and IM criteria. When the information from the response times was leveraged, different weights were applied to modulate the impact of the time/speed leverage. The design variables unique to the VL-CCT were arranged to mimic licensure credential assessments. While in testing, simulees received tests in varying lengths until they are classified into one of the pass/fail groups at 95% confidence. Applying a threshold of $\theta = 0$, if an examinee's interval ability estimate does not include the threshold, testing was terminated and the examinee's class membership was decided based on the location of the estimated ability value, $\widehat{\theta}$. The minimum and maximum test lengths were set at one and 60 noting that TEIs are administered in small numbers along with the extant items.

Evaluation

The testing outcomes were evaluated in four aspects: (i) the final test length, (ii) classification accuracy, (iii) item pool usage, and (iv) testing time. The test lengths were examined by the average across the examinees (i.e., average number of items assigned) and the proportion of examinees that reached the maximum test length (i.e., those that did not meet the termination criterion). The results relating to the classification accuracy were evaluated by four statistics: false positive rate (Type I error), false negative rate (Type II error), sensitivity (correct identification of true pass), and specificity (correct identification of true fail). The item pool usage and testing time efficiency were evaluated using the same criterion statistics as in the preceding study.

Results

Test Length. Table 4 reports average test lengths and proportions of examinees that reached the maximum test length. The conditions that deviate from the optimal selection expectedly required more items to achieve the same confidence level. Compared with MI, the LM selection administered 2.823 more items (η^2 = .779) and showed 1.23% higher rate of hitting the maximum test length (η^2 = .203). Leveraging response times similarly required more items, increasing the test length ($\Delta = 2.581$, $\eta^2 = .482$) and frequency of saturated tests ($\Delta = 2.02\%$, $\eta^2 = .394$). The rate of increase in the test length was largest in MI (Δ = 3.221), followed by IM (2.265) and AM (2.257). The trends across the different time weights showed a similar pattern. The heavier the time weight, the greater the departure from the optimal selection, and thus the longer the tests (η^2 > .200).

Table 4. Test Length in VL-CCT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
Avg	12.48	13.79	14.40	15.14	15.70	15.36	17.15	17.22	17.34	17.62	15.25	17.11	17.10	17.24	17.51
% Max	.101	.110	.115	.120	.123	.113	.126	.129	.129	.133	.112	.126	.127	.129	.131

Note. Avg: Average test length. % Max: Percentage of examinees that received the maximum number of items (i.e., 60). Underlines indicate the best result within each selection method.

The different test lengths administered in each condition causally affect other testing outcomes, including classification accuracy, item pool usage, and testing time. Below, we evaluate these outcomes taking into account the effects of the different test lengths.

Classification Accuracy. Table 5 reports classification outcomes of VL-CCT. The values shown in $\omega = 0$ again suggest that the selection methods performed adequately well. The classification error was constantly small (average = .019; SD = .005) and sensitivity and specificity were maintained competently high (average = .960; SD = .010). Compared with MI, the LM selection showed slightly larger error despite the longer tests ($\eta^2 < .110$). This is because the items selected by the criteria contained less test information than those selected by the MI criterion (e.g., average SE = .383 (MI) versus .426 (LM)). Notwithstanding, the difference in the classification error rates was mostly within the third decimal place (Δ false classification = .003, Δ correct classification = $-.006$), suggesting approximately equal classification performance. The two selection methods in LM performed comparably and showed no significant difference ($p > .450$).

Table 5 also shows that the time leverage generally induced positive outcomes ($\eta^2 < .024$). Although marginal, both the false decision rates and classification accuracy improved as the time/speed was considered in the item selection (Δ false classification = $-.001$, Δ correct classification = .003). The current pattern can be again explained by the different amounts of test information. With the use of more items (2.581 on average), the time-leveraged item selection showed smaller SEs (.343 (time leverage) versus .411 (no time leverage)) and tended to achieve higher classification accuracy. The impact of different time weights was generally impalpable because of the compound effects of the selection method and the test length ($\eta^2 < .008$). Although there were certain weights that led to better or worse classification results, the observed values were mostly comparable and differed only in the third decimal place.

Item Pool Use. Table 6 summarizes results relating to the item pool usage. Location matching again showed significant improvement in the pool usage, yielding substantially smaller χ^2 (Δ = 10.872), lower test overlap rate ($\Delta = -1.73\%$), and smaller percentages of over- and under-used items (Δ Max = -15.07%, Δ Under = $-.79\%$). Recall that LM methods consumed more items, and the improvement in the pool usage may be the result of the longer tests administered. As we closely examine the evaluation statistics, we found that the difference in the statistics was mostly related to the choice of the item selection method (e.g., $\eta^2 = .894$ in χ^2) than to the test length ($\eta^2 = .148$). This means that selecting items on the LM criterion had a more significant direct effect on the even usage of item pools than did the indirect effect that occurs through the test length.

Table 5. Classification Accuracy in VL-CCT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
FP	.018	.017	.017	.018	.018	.020	.019	.019	.019	.018	.020	.018	.019	.019	.018
FN	.018	.017	.017	.018	.019	.022	.019	.019	.019	.019	.021	.019	.019	.019	.018
Sen	.965	.965	.966	.965	.963	.957	.962	.962	.962	.962	.957	.961	.962	.961	.963
Spc	.964	.965	.965	.964	.963	.961	.962	.961	.963	.964	.960	.963	.963	.962	.963

Note. FP: False positive. FN: False negative. Sen: Sensitivity. Spc: Specificity. Underlines indicate the best result within each selection method.

Table 6. Item Pool Usage in VL-CCT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
χ^2	18.97	19.10	18.94	18.58	18.28	8.09	8.58	7.15	6.03	5.24	8.11	8.68	7.18	6.09	5.26
Overlap	.062	.065	.068	.067	.065	.045	.055	.053	.051	.056	.044	.052	.050	.050	.053
% Max	.156	.201	.225	.250	.273	.005	.020	.011	.005	.003	.006	.024	.011	.005	.003
% Under	.019	.017	.017	.019	.017	.011	.008	.004	.003	.002	.011	.008	.004	.003	.002

Note. Overlap: Average test overlap rate between two random examinees. % Max: Percentage of items that reached maximum expose (20% of the sample).% Under: Percentage of under-used items (less than 1%). Underlines indicate the best result within each selection method.

In Table 6, leveraging response times entailed somewhat different results depending on the selection method and the evaluation criterion. In MI, the time-leveraged selection (i.e., $\omega = 1$) led to lower χ^2 ($\Delta = -.690$) but showed higher test overlap rate ($\Delta = .24\%$, $p = .745$) and more frequent item overuse ($\Delta = 11.67\%$). In LM, leveraging working speed improved χ^2 ($\Delta = -2.850$), item over- and under-use (Δ Max $= -.28\%$, $p = .003$; Δ Under $= -.89\%$) but intensified the test overlap ($\Delta = .96\%$, $p = .009$). We find that the distinct patterns in the pool usage are related to the characteristics of the selected items and the test lengths (see Appendix C for illustrative examples). When items were selected through MIT, items with little time intensity were inordinately preferred, frequently using facile items. Interestingly, the use of time leverage helped alleviate the strong preference for the discriminating items and led to better use of low-discriminating items. The net effect in the item over- and under-use made the exposure distribution more uniform-like, consequently improving the skewness. The items selected by the speed-leveraged LM, on the other hand, were dispersed across the wide ranges of item parameter values and tended to exhibit a normal-like exposure distribution. As items were selected to match the examinees' ability and speed levels, the item exposure distribution tended to exhibit a strong central tendency, improving the evaluation statistics on the whole. In both the criteria, the increase in the test overlap rate appeared to be due to the increased use of items. Varying time weights similarly had mixed effects according to the selection method and the evaluation criterion. In MI, the greater time weight helped improve χ^2 ($\eta^2 = .254$) but deteriorated the test overlap rate ($\eta^2 = .000$) and item over- and under-use ($\eta^2_{\text{max}} = .747$; $\eta^2_{\text{under}} = .001$). Increasing speed leverage generally led to more balanced item pool usage ($\eta^2_{\text{chisq}} = .549$; $\eta^2_{\text{max}} = .033$; $\eta^2_{\text{under}} = .265$) while it intensified the test overlap rate ($\eta^2 = .004$).

Testing Time. Table 7 reports descriptive statistics of the testing times observed from VL-CCT. The results indicate that the time leverage entailed different patterns according to the selection method. In MI, considering response times led to shorter testing times ($\Delta = -.528$, $\eta^2 = .818$); in LM, leveraging speed led to longer testing times ($\Delta = .179$, $\eta^2 = .522$). Recall that introducing the time/speed leverage entailed longer tests in both cases. Despite the increase in the test length, MIT showed shorter testing times, whereas the speed-leveraged LM led to longer testing times. The pattern in the MIT was largely due to the intense use of less time-consuming items. MIT had a strong tendency to prefer items with little time intensity and this led to shorter testing times despite the additional item assignments. In LM, items were sampled across a wide range of time intensities, and the administration of more items naturally led to longer testing times. We note that although the speed leverage did not save the average testing times, it still helped regulate the maximum and variance of the testing times in LM (Δ Max $= -.180$, $\eta^2 = .021$; Δ SD $= -.280$,

Table 7. Logarithm of Testing Times in VL-CCT.

	Maximum Information					Average Matching					Intermediate Matching				
ω	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00	.00	.25	.50	.75	1.00
Avg	2.59	2.35	2.16	2.09	2.06	2.82	3.03	3.01	2.99	3.00	2.81	3.02	3.00	2.99	2.99
Max	7.92	7.57	7.30	7.12	7.10	7.93	7.86	7.81	7.75	7.73	7.90	7.90	7.85	7.78	7.75
SD	1.80	1.78	1.75	1.74	1.72	1.78	1.65	1.60	1.55	1.49	1.78	1.66	1.61	1.54	1.50

Note. Avg = Average of log total testing time. Max = Maximum of log testing time. SD = Standard deviation of log testing time. Underlines indicate the best result within each selection method.

$\eta^2 = .900$). The trends across the different time weights were generally consistent with the patterns described above. Increasing the time weight improved all time-wise evaluation statistics in MI as a result of the greater use of facile items ($\eta^2 > .228$). Increasing weight on the speed leverage in LM increased the average testing times ($\eta^2 = .165$) while reducing the maximum and variation ($\eta^2_{\text{max}} = .013$; $\eta^2_{\text{SD}} = .787$).

Conclusion

The purpose of this study was to explore adaptive testing strategies for polytomous TEIs. Current approaches to administering polytomous items are mostly developed based on the information measures and tend to experience severe asymmetry in item usage. If the selection criterion considers item response times, the usage problem can exacerbate in preference for informative and facile items. Recognizing the concern on the skewed use of TEIs, the present study sought to explore alternative item selection methods that can lead to more balanced pool usage. Our approach was to match item locations to examinee's trait levels so that items can be adapted in relative difficulty and time intensity. Since examinees have various trait levels, it was reckoned that location-matched item selection would lead to more diverse item sampling across the wide range of location values and use item pools more exhaustively and evenly.

Numerical experiments from the Monte Carlo simulation suggest that location-matched item selection indeed achieves significantly better and greater balance in item pool usage. The two LM methods, AM and IM, showed distinctly smaller χ^2, lower test overlap rate and smaller percentages of over- and under-used items when compared with MI. Leveraging speed in LM also had clear effects, reducing the average testing times and regulating variation across examinees. The speed leverage also helped improve the item pool usage through additional matching in time intensities.

While the empirical evaluation showed that the location matching and speed leverage deliver significantly better administrative outcomes, it is important to note that they can also entail cost in the measurement outcomes. Both approaches depart from the optimal item selection and can increase measurement error or lose measurement efficiency. For example, when the test length is fastened at a constant value, the measurement cost can appear in the ability estimates and/or standard error. When the test length is allowed to vary, they can lead to longer tests, consequently increasing the test completion times. Our experiment in the two test settings suggests that although the procedures do come with some measurement cost, the magnitude of the loss is generally marginal and is likely to occur in a limited scale in real settings. Operational testing programs typically adopt sufficiently long tests and administer TEIs in small numbers along with the regular items in the existing formats. In such circumstances, the loss in the measurement—either in the precision or test length can be easily offset by the extant items.

We conclude the article with a final remark on the use of timing information in real assessments. Although our simulation study showed that leveraging time information in the item selection does effectively improve testing time efficiency, the application to real settings must carefully consider consequential validity of using timing information. Reliable modeling of behavioral data is often difficult to achieve in real assessments and utilizing auxiliary information during decision-making process can introduce fairness issues. For making capital of the response time information, sufficient research and careful decision process must be preceded. Our work in this study is an attempt to inform such decisions when compromise is needed between the measurement precision and administrational needs.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Council of State Boards of Nursing; (UTA21-000242).

ORCID iD

Hyeon-Ah Kang https://orcid.org/0000-0003-4496-6467

Supplemental Material

Supplemental material for this article is available online

Notes

1. Copious research exists in the measurement literature that imposes heuristic nonstatistical constraints in item selection (e.g., Chang & Ying, 1999; Cheng & Chang, 2009; Kingsbury & Zara, 1989; Sympson & Hetter, 1985; van der Linden, 2000). This study focuses on the statistical selection methods that achieve better item pool usage by their inherent design.
2. The other indices in Ali et al. (2015) that are not pursued in this study are the median- and trimmed-mean-based indices. These indices consider step difficulties around the center categories only and do not accommodate a wide range of ability levels.
3. Minimizing distance between β_j and $\widehat{\tau}$ progressively improves the estimation of τ (i.e., the root of the score function) and helps improve the precision of the item selection.
4. See Appendix B for supplemental simulations under other correlations.
5. Previous studies on polytomous CAT commonly considered 100 or fewer items with four response categories (e.g., Gorin et al., 2005; Lee & Dodd, 2012; Leroux et al., 2019; Pastor et al., 2002; Penfield, 2006). This study assumes a moderate size of 200 items with greater score multiplicity to mimic operational TEIs (e.g., Kang, Han, Betts, & Muntean, 2022; Kang, Han, Kim, & Kao, 2022).
6. EAP was applied on the first item and when the response scores are all minimum or maximums; otherwise MLE was applied.
7. The time-related results were evaluated on the log metric to assess the mean and variation under the approximate normal distribution.

References

Ali, U. S., Chang, H., & Anderson, C. J. (2015). Location indices for ordinal polytomous items based on item response theory. *ETS Research Report Series*, *2015*(2), 1–13, https://doi.org/10.1002/ets2.12065

Betts, J., Muntean, W., Kim, D., & Kao, S. (2021). Evaluating different scoring methods for multiple response items providing partial credit. *Educational and Psychological Measurement*, *82*(1), 151–176. https://doi.org/10.1177/0013164421994636

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(Pt 2), 369–383. https://doi.org/10.1348/000711008X304376

Cheng, Y., Diao, Q., & Behrens, J. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, *49*(2), 502–512. https://doi.org/10.3758/s13428-016-0712-6

Choe, E., Kern, J. L., & Chang, H.-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*(2), 135–158. https://doi.org/10.3102/1076998617723642

Cohen, J. (1988). *Statistical power analysis for the social sciences* (2nd. edition). Lawrence Erlbaum Associates.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670. https://doi.org/10.3102/1076998611422912

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement*, *29*(6), 433–456. https://doi.org/10.1177/0146621605280072

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement*, *72*(3), 493–509. https://doi.org/10.1177/0013164411422903

Kang, H.-A., Han, S., Betts, J., & Muntean, W. (2022). Computerized adaptive testing for testlet-based innovative items. *British Journal of Mathematical and Statistical Psychology*, *75*(1), 136–157. https://doi.org/10.1111/bmsp.12252

Kang, H.-A., Han, S., Kim, D., & Kao, S.-C. (2022). Polytomous testlet response models for technology-enhanced innovative items: Implications on model fit and trait inference. *Educational and Psychological Measurement*, *82*(4), 811–838. https://doi.org/10.1177/00131644211032261

Kang, H.-A., Zheng, Y., & Chang, H.-H. (2021). Online calibration of a joint model of item responses and response times in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *45*(2), 175–208. https://doi.org/10.3102/1076998619879040

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54–75. https://doi.org/10.1037/a0014877

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159–175. https://doi.org/10.1177/0013164411411296

Leroux, A. J., Waid-Ebbs, J. K., Wen, P.-S., Helmer, D. A., Graham, D. P., O'Connor, M. K., & Ray, K. (2019). An investigation of exposure control methods with variable-length CAT using the partial credit model. *Applied Psychological Measurement*, *43*(8), 624–638. https://doi.org/10.1177/0146621618824856

Lord, M. F. (1970). Some test theory for tailored testing. In W. H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (p. 139–183). Harper & Row.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), 159–176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Pastor, D. A., Dodd, B. G., & Chang, H.-H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, *26*(2), 147–163. https://doi.org/10.1177/01421602026002003

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education*, *19*(1), 1–20. https://doi.org/10.1207/s15324818ame1901_1

Sun, S.-S., Tao, J., Chang, H.-H., & Shi, N.-Z. (2012). Weighted maximum-a-posteriori estimation in tests composed of dichotomous and polytomous items. *Applied Psychological Measurement*, *36*(4), 271–290. https://doi.org/10.1177/0146621612446937

Sympson, J. B., & Hetter, R. D. (1985, October). 21-25 Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the military testing association* (pp. 973–977). Military Testing Association.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. https://doi.org/10.1177/0146621609349800

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–53). Kluwer Academic Publishers.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139. https://doi.org/10.1007/s11336-009-9129-9

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. https://doi.org/10.1177/01466219922031329

van der Linden, W. J., & van Krimpen-Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265. https://doi.org/10.1007/bf02294800

van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *26*(4), 393–411. https://doi.org/10.1177/014662102237796

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 207–214). Springer-Verlag.