the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 109–141

www.wileyonlinelibrary.com

Utilizing response times in cognitive diagnostic computerized adaptive testing under the higher-order deterministic input, noisy 'and' gate model

Hung-Yu Huang*
Department of Psychology and Counseling, University of Taipei, Taiwan

Methods of cognitive diagnostic computerized adaptive testing (CD-CAT) under higher-order cognitive diagnosis models have been developed to simultaneously provide estimates of the attribute mastery statuses of examinees for formative assessment and estimates of a latent continuous trait for overall summative evaluation. In a typical CD-CAT environment, examinees are often subject to a time limit, and the examinees' response times (RTs) for specific test items can be routinely recorded by custom-made programs. Because examinees are individually administered tailored sets of test items from the item pool, they may experience different levels of speededness during testing and different levels of risk of running out of time. In this study, RTs were considered during the item-selection procedure to control the test speededness and the RTs were treated as useful information for improving latent trait estimation in CD-CAT under the higher-order deterministic input, noisy 'and' gate (DINA) model. A modified posterior-weighted Kullback–Leibler (PWKL) method that maximizes the item information per time unit and a shadow-test method that assembles a provisional test subject to a specified time constraint were developed. Two simulation studies were conducted to assess the effects of the proposed methods on the quality of CD-CAT for fixed- and variable-length exams. The results show that, compared with the traditional PWKL method, the proposed methods preserve a lower risk of running out of time while ensuring satisfactory attribute estimation and providing more accurate estimates of the latent trait and speed parameters. Finally, several suggestions for future research are proposed.

1. Introduction

In computerized adaptive testing (CAT), each examinee is administered a tailored set of test items that are selected based on that examinee's responses to previously administered items, thereby allowing examinees' abilities to be estimated with a higher degree of precision when compared with non-adaptive testing (van der Linden & Glas, 2010). In a typical CAT exam, the examinees' ability levels are scored on a latent continuous scale using item response theory (IRT) models, and one (or more) summative score for a single (or multiple) broadly defined latent trait(s) is calculated to evaluate the examinees' performance. In some cases, for example, in classroom instruction, the mastery statuses of examinees for a given set of skills may be informative and important. Various cognitive diagnosis models (CDMs) have also been developed to classify examinees' mastery levels

*Correspondence should be addressed to Hung-Yu Huang, Department of Psychology and Counseling, University of Taipei, No. 1, Ai-Guo West Road, Taipei 10048, Taiwan (email: hyhuang@go.utaipei.edu.tw).

DOI:10.1111/bmsp.12160

in terms of a set of multidimensional latent binary variables (also called latent attributes); these models collect sufficient diagnostic information to allow instructional interventions to be immediately and efficiently provided to examinees (Rupp, Templin, & Henson, 2010). Based on well-established CAT algorithms and CDMs, methods of cognitive diagnostic CAT (CD-CAT) have been developed to efficiently provide profile information on fine-grained cognitive attributes by tailoring tests to each individual examinee (Cheng, 2009; Hsu, Wang, & Chen, 2013; McGlohen & Chang, 2008).

Both IRT-based CAT and CD-CAT are subject to certain limitations in practical applications. For example, CD-CAT is devoted to ensuring classification accuracy for latent attributes to facilitate formative assessment, whereas IRT-based CAT focuses on the precise evaluation of one or more latent traits. Thus, IRT-based CAT does not permit the assessment of examinees' mastery statuses for a set of specific attributes or skills in the absence of some means of subjective judgment, such as computerized classification testing (Thompson, 2009). To address the practical demands of real testing situations, higher-order CDMs have been proposed (de la Torre & Douglas, 2004) that combine the advantages of latent trait estimation and latent attribute classification; in these models, a second-order continuous latent trait is assumed to govern the mastery of the first-order latent attributes, and the item responses are determined by combinations of the latent attributes. Recently, Hsu and Wang (2015) developed CAT algorithms for the higher-order deterministic input, noisy 'and' gate (DINA) model, using three types of minimum-precision termination rules (i.e., algorithms for variable-length CD-CAT) based on attribute classification precision, latent trait estimation precision, and a combination of both types of precision. They found that the combined criterion performed better than the other two approaches in both latent trait estimation and latent attribute estimation.

Although higher-order CDMs can enable both types of assessment, the second-order continuous latent trait cannot be precisely estimated because of its greater root mean square error (RMSE) of estimation (Hsu & Wang, 2015). This is because the number of latent attributes is often small (e.g., fewer than 10), and the information gained from the latent attribute estimation is insufficient to enable the precise estimation of the second-order latent trait. Increasing the number of latent attributes may improve the latent trait estimation, but this approach is not practical and can result in higher estimation errors for the latent attributes (Cheng, 2009). Therefore, an additional source of information is needed if we wish to improve the measurement precision for the second-order latent trait in CD-CAT under higher-order CDMs.

In a typical CD-CAT environment, examinees are often subject to a time limit; for example, an operational CD-CAT exam of a fixed-length of 36 items for English achievement was administered to students in Grades 5 and 6 in China with a time limit of 40 min (Liu, You, Wang, Ding, & Chang, 2013). Time limits are a practical necessity in the administration of exams, and test speededness should be considered in most tests (Shao, Li, & Cheng, 2016). In CAT or CD-CAT, the response times (RTs) of each test taker are often routinely recorded and can be used to monitor the amount of time that an examinee spends in responding to each item. The use of RTs to improve the efficiency of CD-CAT is justified because a short test duration with sufficient test items is desirable for obtaining precise estimates of examinees' abilities while simultaneously maintaining the respondents' motivation and ensuring that the resources (e.g., computer stations) available for a particular testing window can be managed appropriately by limiting the percentage of examinees who will exceed the pre-specified time limits (Sie, Finkelman, Riley, & Smits, 2015). Several procedures have been developed for speededness control or improved item selection in the CAT context (Fan, Wang, Chang, & Douglas, 2012; van der Linden,

2008, 2009a; van der Linden & Xiong, 2013). Two major control procedures of this kind, which were adopted in this study, merit further discussion. The first is the shadow-test approach (STA), in which a shadow test is assembled by selecting the items that are maximally informative, subject to the time limit specified for the test (van der Linden, 2009a; van der Linden & Xiong, 2013). The second is a new item-selection method introduced by Fan *et al.* (2012) for choosing the candidate items that yield the maximum information per time unit; these authors found that this modified item-selection method is operationally easier than other control methods and can reduce testing time while maintaining acceptable measurement precision.

In addition to facilitating the consideration of practical constraints, RTs can be used to improve ability estimation based on the joint distribution of the ability and speed parameters in a hierarchical modelling framework (van der Linden, 2008, 2009b; van der Linden, Entink, & Fox, 2010). Utilizing such joint distributions in CD-CAT under higher-order CDMs may improve latent trait estimation by virtue of the additional information provided by the RT data. Imprecise latent trait estimation not only threatens the validity of summative assessments using higher-order CDMs but also compromises the termination criterion for variable-length CD-CAT.

Past studies have generated abundant results on improving person- and item-parameter estimation and on making CAT more efficient by using RT information (van der Linden, 2008, 2009a; van der Linden *et al.*, 2010). However, these studies have focused on the estimation of broadly defined latent traits and have paid little attention to classification in a set of cognitive diagnostic attributes as RT data have been collected. To the best of our knowledge, few studies have investigated the effects of using RT information on speededness control and on attribute and latent trait estimation in CD-CAT. Inspired by previous research, this study aimed to develop a new set of CD-CAT algorithms for efficient speededness control and precise person–parameter estimation using a heuristic approach that considers the maximum information per item and an assembling approach that implements a shadow test, which has never previously been used in CD-CAT. In addition, the procedure for controlling the percentage of examinees who exceed the acceptable range of the given time limit, as developed in the literature, was first applied to CD-CAT for the exam termination criterion in this study. Furthermore, joint hierarchical modelling of item responses and RTs can be used to associate the speed parameter with a latent continuous latent in the higher-order CDMs or to associate the speed parameter with the latent attributes in the conventional CDMs, which verify the applicability and flexibility of the new CD-CAT algorithms. Therefore, evaluating the effectiveness of the developed CD-CAT algorithms on speededness control and person–parameter estimation in fixed- and variable-length CD-CAT serves as the major purpose of the current study.

Although the idea of using RT data to improve the efficiency of CAT by maximizing information per item, and assembling a shadow test is not completely novel in nature, the two approaches for controlling speededness were developed individually in the literature on IRT-based CAT, and their effects on speededness control are never systematically compared in the CD-CAT condition until this study. CD-CAT has great potential as an instructional remedy in the classroom, and preventing the unnecessary loss of lecture time is always desirable for embedded or formative assessment (DiBello & Stout, 2007; Wang, 2013). Thus, the simultaneous consideration of precise attribute classification and time limits is particularly important and should be evaluated for its effectiveness via systematic simulation, which serves as a practical contribution of this study.

This study is organized as follows. First, I briefly review the higher-order DINA model and the RT model. Second, CD-CAT algorithms based on the higher-order DINA model are

introduced for fixed- and variable-length exam conditions, using the posterior-weighted Kullback–Leibler (PWKL) information method (Cheng, 2009) as the item-selection method. Third, I introduce the modified item-selection methods and the time-limited stopping rule used in this study to consider the imposition of a time limit on a test. Fourth, I report a series of simulation studies conducted to evaluate the effectiveness of the new CD-CAT algorithms when using a hierarchical joint prior distribution with respect to the latent trait and speed parameters in the higher-order DINA model and to the latent attributes and speed parameters in the conventional DINA model. Finally, I draw several conclusions from the results and present suggestions for future studies.

2. The higher-order DINA model with the use of RTs

In a CDM, a vector of latent binary attributes is used to represent an examinee's mastery status in each attribute; this vector is written as $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$, where $\alpha_{ik} = 1$ indicates that examinee i has mastered attribute k and $\alpha_{ik} = 0$ indicates otherwise. In addition, a **Q**-matrix that represents the item-to-attribute mapping is specified as a $J \times K$ matrix, in which entry q_{jk} is equal to 1 if attribute k is required to provide the correct response to item j and is equal to 0 otherwise. Classified as a non-compensatory model, the DINA model partitions the 2^K possible latent attribute vectors (or latent classes) into two latent groups, where $\xi_{ij} = 1$ indicates that examinee i possesses all the attributes required to solve item j and $\xi_{ij} = 0$ indicates otherwise. Therefore, in the DINA model, each examinee i with attribute vectors in the same group is assumed to have the same probability of answering item j correctly, which is defined as

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_j)^{\xi_{ij}} g_j{}^{(1-\xi_{ij})}, \tag{1}$$

such that

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{2}$$

Here, Y_{ij} is the response of examinee i to item j (a value of 1 indicates the correct response, whereas 0 indicates an incorrect response), s_j is the probability of an incorrect response to item j when all attributes required for item j have been mastered (i.e., the slip parameter) and g_j is the probability of a correct response to item j when a person lacks at least one of the attributes required for that item (i.e., the guessing parameter).

Because the latent attributes are seldom independent, a hierarchical framework can be established in which a second-order continuous latent trait is assumed to govern the mastery statuses of the first-order latent attributes, such that an examinee with a higher value of this latent trait is expected to have a higher probability of mastering attributes. The higher-order DINA model was developed to describe the acquisition of attributes by constraining the probability of possessing each attribute to be a function of a broadly defined latent trait θ (de la Torre & Douglas, 2004) as follows:

$$P(\alpha_{ik} = 1|\theta_i) = \frac{\exp[\lambda_{1k}(\theta_i - \lambda_{0k})]}{1 + \exp[\lambda_{1k}(\theta_i - \lambda_{0k})]}. \tag{3}$$

Here, λ_{0k} and λ_{1k} are the location and discrimination parameters, respectively, for attribute k and θ_i is the general ability level of examinee i, which is assumed to follow a

standard normal distribution. Equation (3) uses a logit link function and can be replaced by the probit link function using the normal ogive modelling framework, and the results of the logit link can be approximated by those of the probit link when multiplied by a constant of 1.7. Because the logit link function is easier to implement than the probit link function, the logistic formulation in equation (3) was adopted in this study.

When a test is administered on a computer, the RTs are usually routinely recorded. RT data can be described by a lognormal model (van der Linden, 2006). Let t_{ij} denote the RT for examinee i on item j, and let this quantity have the following probability density function:

$$f(t_{ij};\tau_i,\eta_j,\delta_j)=\frac{\eta_j}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\eta_j(\ln t_{ij}-(\delta_j-\tau_i))\right]^2\right\}. \tag{4}$$

Here, τ_i is the speed parameter for examinee i, η_j is the discrimination parameter of item j with respect to speed and δ_j is the time-intensity parameter of item j. When examinees' responses to test items are modelled by a specific response model (e.g., IRT models), the latent trait parameter underlying the item responses and the speed parameter influencing the RTs are assumed to be jointly normally distributed at the second level. Also, separate measurement models for accuracy and speed are treated as the first-level components to capture all systematic variation in the item responses and RTs of examinees using a hierarchical modelling approach (van der Linden, 2007). Compared to the traditional IRT models, hierarchical modelling of the ability and speed parameters facilitates the use of RTs as collateral information in the estimation of the IRT parameters and improves estimation efficiency (van der Linden *et al.*, 2010).

The hierarchical framework is very flexible because the incorporation of any measurement model for either accuracy or speed is allowed by taking the possible dependences to a second level of modelling to satisfy the conditional independence assumption (Entink, Kuhn, Hornke, & Fox, 2009b, p. 62). In fact, this framework is not limited to any specific response or RT model; consequently, other models could be substituted for the IRT or lognormal RT model (van der Linden, 2009b, pp. 264–265; van der Linden *et al.*, 2010, p. 330). Therefore, it is justifiable to assume that the latent trait parameter θ of the higher-order DINA model and the speed parameter τ of the RT model follow joint distributions in hierarchical modelling, which can be expressed as

$$(\theta,\tau)\sim N_2(\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{5}$$

where the mean vector $\boldsymbol{\mu}$ is constrained by the zero-mean-vector condition for model identification, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix. Thus, the correlation between θ and τ is expected to extract additional information to improve the estimation of examinees' abilities, as in IRT models (Entink *et al.*, 2009b; van der Linden *et al.*, 2010).

Because the θ and τ parameters are allowed to correlate in the hierarchical model, we can express the relationship by regressing the latent trait parameter on the speed parameter as

$$\theta_i=\beta\tau_i+\varepsilon_i, \tag{6}$$

where β is the regression coefficient, equivalent to the correlation coefficient when both the θ and τ parameters follow a standard normal distribution, and ε_i is a residual term assumed to be normally distributed. Note that because the θ parameter is constrained to

follow a normal distribution with a mean of zero and a variance of one for model identification, the intercept term is not estimated in equation (6). Substituting equation (6) for the θ parameter in equation (3) leads to the following:

$$P(\alpha_{ik} = 1|\theta_i) = \frac{\exp[\lambda_{1k}(\beta\tau_i + \varepsilon_i - \lambda_{0k})]}{1 + \exp[\lambda_{1k}(\beta\tau_i + \varepsilon_i - \lambda_{0k})]}. \quad (7)$$

When appropriate, a quadratic regression of the θ parameter on the τ parameter can be modelled to represent the non-linear relationship between the θ and τ parameters, for example, in some circumstances of attitude and personality assessments (Molenaar, Tuerlinckx, & van der Maas, 2015).

The parameters in the higher-order DINA model and the lognormal RT model can be calibrated jointly using Bayesian estimation. Let ω be a parameter to be estimated, let $p(\omega)$ be the prior distribution, and let $\mathbf{Y}$ be the observed responses. The posterior density of $\omega|\mathbf{Y}$ can be denoted as

$$p(\omega|\mathbf{Y}) = \frac{f(\mathbf{Y}|\omega)p(\omega)}{m(\mathbf{Y})}, \quad (8)$$

where $m(\mathbf{Y})$ is the marginal distribution of $\mathbf{Y}$ (Casella & Berger, 2002). When multiple parameters are involved and many random-effect variables must be estimated, Markov chain Monte Carlo (MCMC) methods can be used to simulate the joint posterior distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters of interest (Entink, Fox, & van der Linden, 2009a; Entink *et al.*, 2009b). Alternatively, the generalized linear factor model can be used to fit the item response and RT data in the hierarchical model using marginal maximum likelihood (MML) estimation by restricting the item parameters as fixed effects; many well-established and user-friendly computer programs (e.g., Mplus) are readily available to ordinary users, in contrast to what would be required by the complex mathematical nature of the Bayesian sampling methodology. Readers who are interested in a comprehensive comparison of the Bayesian and MML estimation methods in hierarchical models can refer to the study by Molenaar *et al.* (2015).

Recently, Zhan, Jiao, and Liao (2018) implemented the Gibbs sampler of the MCMC methods in the JAGS computer program to estimate the model parameters in the higher-order DINA model with the use of RTs. They found that the attribute classification accuracy and the precision of model parameter estimation can be improved when simultaneously modelling the association of the second-order latent trait parameter with the speed parameter. Supported by the evidence found in the work of Zhan *et al.* (2018), which included simulation and empirical studies, when ideal testing situations are satisfied (e.g., sufficient calibration sample size and model data goodness of fit), the model parameters of the higher-order DINA model and the lognormal RT model in the item pool in CD-CAT can be readily calibrated for further administration.

3. Directly modelling the association of the latent attributes with the speed parameter

To facilitate the classification accuracy of the latent attribute and to maximize the efficiency of CD-CAT with the use of RT information, in addition to constructing a

hierarchal model of the second-order continuous latent ability parameter (θ) and the speed parameter (τ), as described above, another approach can be considered to directly specify the relationship between attribute mastery and the speed parameter rather than using the θ parameter to affect the chance of possessing attributes. The multivariate Bernoulli distribution (MBD) assumes K attributes follow a K-dimensional correlated Bernoulli distribution (Dai, Ding, & Wahba, 2013), and it can be viewed as an alternative way to model the relationship of $\boldsymbol{\alpha}$ and τ. Wang and Qiu (2018) developed multilevel CDMs using the MBD approach to incorporate the observable person covariates (e.g., gender) to predict the mastery status of latent attributes. However, in the present study, the latent speed parameter is used as a covariate instead of an observable variable; therefore, the relationship of $\boldsymbol{\alpha}$ and τ are simultaneously considered.

Because the α parameter is a categorical latent variable and the τ parameter is continuous, a logistic regression is used to present the probability of mastering attribute k for examinee i given the predictor τ, which is given by

$$P(\alpha_{ik} = 1|\gamma_{0k}, \gamma_{1k}, \tau_i) = \frac{\exp(\gamma_{0k} + \gamma_{1k}\tau_i)}{1 + \exp(\gamma_{0k} + \gamma_{1k}\tau_i)}, \tag{9}$$

where γ_{0k} and γ_{1k} are the intercept term and regression weight, respectively, for attribute k. The joint likelihood function of the DINA model using the MBD approach can thus be obtained by combining equations (1), (2) and (9) as follows:

$$L(\mathbf{Y}|\mathbf{s}, \mathbf{g}, \boldsymbol{\alpha}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\left[(1 - s_j)^{Y_{ij}} s_j^{1-Y_{ij}}\right]^{\xi_{ij}}\left[g_j^{Y_{ij}}(1 - g_j)^{1-Y_{ij}}\right]^{1-\xi_{ij}}. \tag{10}$$

Let $\boldsymbol{\alpha}_c$ denote a specific attribute pattern, and let γ_0 and γ_1 denote the vector for the intercept and regression weight parameters, respectively. Given τ, the predicted probability of classifying in latent class $\boldsymbol{\alpha}_{c(i)}$ for examinee i can be expressed as

$$P(\boldsymbol{\alpha}_{c(i)}|\gamma_0, \gamma_1, \tau_i) = \prod_{k=1}^{K} P(\alpha_{ik} = 1)^{\zeta_k} P(\alpha_{ik} = 0)^{1-\zeta_k}, \tag{11}$$

where ζ_k is an indicator and is equal to one if attribute k is present in $\boldsymbol{\alpha}_c$, and zero otherwise. The probability of $\boldsymbol{\alpha}_c$ predicted by the τ variable can be used as empirically individualized prior information for the adaptive test and it can be expected to improve the attribute classifications of examinees in CD-CAT. As in the item pool calibration, the intercept and regression weight parameters can be estimated from pre-test data using MML or Bayesian estimation.

4. CD-CAT algorithms

4.1. Item-selection procedure in CD-CAT

When applying the higher-order DINA model in CD-CAT, it is necessary to choose an item-selection procedure, methods of attribute and ability estimation, and a termination rule. Several item-selection procedures for CD-CAT have been proposed for the adaptive administration of test items to examinees (Cheng, 2009; Kaplan, de la Torre, & Barrada,

2015; McGlohen & Chang, 2008; Wang, 2013). In higher-order CD-CAT, both the latent attributes and the latent trait should be estimated, and information about both $\boldsymbol{\alpha}$ and θ can be incorporated into the item-selection process (Wang, Chang, & Douglas, 2012). However, a dual-calibration process, which is a process in which different sets of item parameters are calibrated based on different models (an IRT model and a CDM), and the use of different sets of item parameters for item selection are not common and should be adopted with caution. This is because the underlying latent constructs are dramatically different in nature between the two types of models, and an IRT model and a CDM might not yield similar fits unless the latent attributes exhibit a linear hierarchy (Hsu & Wang, 2015). Therefore, I adopted a single-calibration process as the item-selection method in this study. I chose the PWKL procedure (Cheng, 2009), which is a popular and commonly used item-selection method with respect to the latent attributes in the literature. In addition, the PWKL method has been found to perform more satisfactorily than other item-selection methods, such as the Kullback–Leibler (KL) information and Shannon entropy methods, in terms of the efficiency of classification of examinees' attributes (Cheng, 2009), and it has been used in many previous studies (Cheng, 2009; Hsu & Wang, 2015; Hsu *et al.*, 2013; Mao & Xin, 2013). However, in some situations, the opposite effects were observed for these item-selection methods. For example, when a short test was used (fewer than 10 items), the Shannon entropy method and its modified approach of the mutual information method were found to be more efficient than the PWKL method in terms of the attribute mastery classification in CD-CAT (Wang, 2013; Zheng & Chang, 2016).

The PWKL method was derived based on the KL information measure to quantify the distance or discrepancy between two probability distributions. Let $P(Y_{ij} = y|\hat{\boldsymbol{\alpha}}_i)$ and $P(Y_{ij} = y|\boldsymbol{\alpha}_i)$ be the conditional distributions of the response of examinee i to item j given the current estimated latent class $\hat{\boldsymbol{\alpha}}_i$ and given the true latent class $\boldsymbol{\alpha}_i$ respectively, for examinee i. By multiplying it by the corresponding posterior distribution of a latent class c for examinee i given that examinee's responses to $(n-1)$ items, the KL information measure can be transformed into the PWKL information measure, which is expressed as

$$\text{PWKL}_j(\hat{\boldsymbol{\alpha}}_i) = \sum_{c=1}^{2^K} \left\{ \left[\sum_{y=0}^{1} \log \left(\frac{P(Y_{ij} = y|\hat{\boldsymbol{\alpha}}_i)}{P(Y_{ij} = y|\boldsymbol{\alpha}_c)} \right) P(Y_{ij} = y|\hat{\boldsymbol{\alpha}}_i) \right] P(\boldsymbol{\alpha}_c|\mathbf{y}_{i,n-1}) \right\}. \tag{12}$$

The posterior distribution of latent class c for examinee i is

$$P(\boldsymbol{\alpha}_c|\mathbf{y}_{i,n-1}) = \frac{P(\boldsymbol{\alpha}_c) L(\mathbf{y}_{i,n-1}|\boldsymbol{\alpha}_c)}{\sum_{c=1}^{2^K} P(\boldsymbol{\alpha}_c) L(\mathbf{y}_{i,n-1}|\boldsymbol{\alpha}_c)}, \tag{13}$$

where $P(\boldsymbol{\alpha}_c)$ is the prior distribution for latent class c that is defined to follow a uniform distribution and $L(\mathbf{y}_{i,n-1}|\boldsymbol{\alpha}_c)$ is the likelihood function of the responses to the $(n-1)$ items for examinee i.

4.2. Methods of attribute classification and ability estimation in CD-CAT

When an examinee completes a set of test items in CD-CAT, the mastery state (i.e., attribute profile or latent class) of that examinee for the latent attributes of interest is

estimated by comparing the examinee's posterior probabilities of belonging to each possible latent class c according to equation (13). The latent class with the largest posterior probability is selected as the attribute profile estimate. In addition, these posterior probabilities are used as a basis for the termination criterion in variable-length CD-CAT because the measurement precision of latent attribute classification is strongly related to the posterior probability (Hsu *et al.*, 2013).

For CD-CAT under the higher-order DINA model, in addition to the attribute classification, the second-order latent trait should be estimated. Assuming local independence, given a vector of attribute estimates $\hat{\boldsymbol{\alpha}}$, the likelihood function across all possible latent classes can be presented as

$$L(\hat{\boldsymbol{\alpha}}|\theta) = \sum_{c=1}^{2^K} \left[\prod_{k=1}^{K} P(\hat{\alpha}_{ck}|\theta)^{\hat{\alpha}_{ck}} \times Q(\hat{\alpha}_{ck}|\theta)^{(1-\hat{\alpha}_{ck})} \right] P(\hat{\boldsymbol{\alpha}}_c), \tag{14}$$

where $P(\hat{\alpha}_{ck}|\theta)$ and $Q(\hat{\alpha}_{ck}|\theta)$ are the probabilities of mastering and not mastering attribute k, respectively, in latent class c, according to equation (3), and $P(\hat{\boldsymbol{\alpha}}_c)$ is the posterior probability of latent class c as defined in equation (13). When appropriate, a prior normal distribution $f(\theta)$ with a mean of zero and a variance of one can be assumed for θ, and the posterior distribution of θ can be expressed as

$$g(\theta|\hat{\boldsymbol{\alpha}}) \propto L(\hat{\boldsymbol{\alpha}}|\theta) \times f(\theta). \tag{15}$$

The Bayesian expected *a posteriori* estimate for θ and the standard error of $\hat{\theta}$ can be obtained by integrating $g(\theta|\hat{\boldsymbol{\alpha}})$ with respect to θ; this approach was used in the work of Hsu and Wang (2015) and was also adopted in this study.

4.3. Termination rules in variable-length CD-CAT

In the higher-order CDM framework, both the second-order latent trait (θ) and the multiple binary attributes ($\boldsymbol{\alpha}$) are of importance and interest because the former can be viewed as an assessment of general aptitude, and the latter, as useful diagnostic information for evaluating specific knowledge through a parsimonious modelling approach (de la Torre & Douglas, 2004). Therefore, when building a variable-length CD-CAT system under the higher-order DINA model, the precision of the measurement of attribute classification and latent trait estimation should be considered simultaneously. Relying solely on the precision of attribute estimation to set the termination rule does not guarantee that the specified level of precision can be achieved for the second-order latent trait estimation. In their pioneering study, Hsu and Wang (2015) showed that the RMSEs of second-order latent trait estimates when considering both the precisions of $\hat{\boldsymbol{\alpha}}$ and $\hat{\theta}$ were always smaller than those when considering only the precision of $\hat{\boldsymbol{\alpha}}$ (Hsu & Wang, 2015, pp. 135–138). Based on evidence supported in the literature, both termination rules for $\hat{\boldsymbol{\alpha}}$ and $\hat{\theta}$ were adopted in this study.

As described above, two sets of termination rules must be specified simultaneously in variable-length CD-CAT under the higher-order DINA model. For the measurement precision control of $\hat{\boldsymbol{\alpha}}$, the maximum posterior probability $P(\hat{\boldsymbol{\alpha}}_{\max})$ can be pre-specified as the minimum-precision criterion. When the posterior probability of the latent class with the largest posterior probability is greater than this threshold value, the CD-CAT process terminates. The measurement precision control procedure of $\hat{\theta}$ is more complicated than

that of $\hat{\boldsymbol{\alpha}}$. As in IRT-based variable-length CAT conditions, a minimum standard error value (i.e., $SE_{\hat{\theta}}$) can be pre-specified for the second-order latent trait estimate in CD-CAT under the higher-order DINA model. When the standard error of the latent trait estimate is smaller than this threshold, CD-CAT terminates.

Because the precisions of $\boldsymbol{\alpha}$ and θ are mutually dependent (Hsu & Wang, 2015, p. 130), the choice of $SE_{\hat{\theta}}$ should be based on the target posterior probability of latent class c that we want to achieve, which is denoted by $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$. Note that $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ does not necessarily have to be equal to $P(\hat{\boldsymbol{\alpha}}_{\max})$, and the target latent class might vary as CD-CAT proceeds. The posterior probability of the target latent class is used to determine the precision level of $\hat{\theta}$, but the maximum posterior probability is used to determine the precision level of $\hat{\boldsymbol{\alpha}}$. Both rules are designed for the purpose of terminating the CD-CAT exam. The latent class with the largest posterior probability (see equation [13]) after administering an item in a CD-CAT exam can be chosen as the target latent class (denoted as c) and specified a target posterior probability for this latent class. Based on a series of simulations, Hsu and Wang (2015) recommended that $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ should be higher than $P(\hat{\boldsymbol{\alpha}}_{\max})$. For example, in a four-attribute simulated bank, they found that the use of $SE_{\hat{\theta}}$, when $SE_{\hat{\theta}}$ was set based on $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ larger than $P(\hat{\boldsymbol{\alpha}}_{\max})$, gave a better performance than the use of $SE_{\hat{\theta}}$ when $SE_{\hat{\theta}}$ was set based on $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ equal to $P(\hat{\boldsymbol{\alpha}}_{\max})$ in all manipulated conditions in terms of attribute classification accuracy and latent trait estimation precision (Hsu & Wang, 2015, p. 137).

When $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ is specified for target latent class c, all but latent class c can be assumed to share the remaining posterior probability equally; that is, each latent class c' ($c' \neq c$ represents the remaining latent classes) is expected to have a posterior probability of $P_{\theta}(\hat{\boldsymbol{\alpha}}_{c'}) = [1 - P_{\theta}(\hat{\boldsymbol{\alpha}}_c)]/(2^K - 1)$. Accordingly, the minimum standard error of $\hat{\theta}$ conditional on $P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ can be expressed as

$$SE_{\hat{\theta}|P_{\theta}(\hat{\boldsymbol{\alpha}}_c)} = \sqrt{\int (\theta - \hat{\theta})^2 \times \frac{L_{\theta}(\hat{\boldsymbol{\alpha}}|\theta) \times f(\theta)}{\int L_{\theta}(\hat{\boldsymbol{\alpha}}|\theta) \times f(\theta) d\theta} d\theta}. \tag{16}$$

The likelihood function across all possible latent classes is

$$\begin{aligned} L_{\theta}(\hat{\boldsymbol{\alpha}}|\theta) &= \prod_{k=1}^{K} P(\hat{\alpha}_{ck}|\theta)^{\hat{\alpha}_{ck}} \times Q(\hat{\alpha}_{ck}|\theta)^{(1-\hat{\alpha}_{ck})} P_{\theta}(\hat{\boldsymbol{\alpha}}_c) \\ &+ \sum_{c' \neq c} \left[\prod_{k=1}^{K} P(\hat{\alpha}_{c'k}|\theta)^{\hat{\alpha}_{c'k}} \times Q(\hat{\alpha}_{c'k}|\theta)^{(1-\hat{\alpha}_{c'k})} \right] P_{\theta}(\hat{\boldsymbol{\alpha}}_{c'}), \end{aligned} \tag{17}$$

where $P(\hat{\alpha}_{ck}|\theta)$ and $P(\hat{\alpha}_{c'k}|\theta)$ are defined as in equation (3). Note that $Q(\hat{\alpha}_{ck}|\theta) = 1 - P(\hat{\alpha}_{ck}|\theta)$, $Q(\hat{\alpha}_{c'k}|\theta) = 1 - P(\hat{\alpha}_{c'k}|\theta)$, and $f(\theta)$ is the prior distribution, which is assumed to follow a standard normal distribution. Therefore, in CD-CAT under the higher-order DINA model, a minimum-precision termination rule can be specified to stop an exam if both of the following conditions are satisfied: the largest posterior probability exceeds $P(\hat{\boldsymbol{\alpha}}_{\max})$, and $SE_{\hat{\theta}}$ is smaller than $SE_{\hat{\theta}}|P_{\theta}(\hat{\boldsymbol{\alpha}}_c)$ (Hsu & Wang, 2015).

5. Incorporating RT information into CD-CAT

In CD-CAT, examinees' attribute profiles are updated after they have answered each administered item. When RT data are collected and the parameters of the RT model are

calibrated in advance, such as during the construction of the item pool, the speed parameter τ_i for examinee i can be estimated after the examinee has completed $(n-1)$ items (van der Linden, 2006, 2008); this estimate is expressed as

$$\hat{\tau}_i = \frac{\sum_{j=1}^{n-1} \eta_j^2(\delta_j - \ln t_{ij})}{\sum_{j=1}^{J} \eta_j^2}. \tag{18}$$

When an examinee receives a new item selected by the CAT system, the time that examinee i is expected to spend answering item j given the interim speed parameter estimate $\hat{\tau}_i$ can be calculated as (Fan *et al.*, 2012)

$$E(t_j|\hat{\tau}_i) = \exp(\delta_j - \hat{\tau}_i + \frac{1}{2\eta_j^2}). \tag{19}$$

Considering this expected time, the PWKL information measure can be modified to choose the next item (denoted by j_n) from among the set of the remaining items (denoted by R_n and containing $J - n + 1$ items) that will yield the maximum information per time unit, as

$$j_n = \arg_{j \in R_n} \max \frac{\text{PWKL}_j(\hat{\boldsymbol{\alpha}}_i)}{E(t_j|\hat{\tau}_i)}. \tag{20}$$

Previous studies have shown that using the expected time to adjust the information measure for the candidate items in CAT can substantially reduce the average time required by an examinee to complete a CAT exam, while preserving an acceptable measurement precision for latent trait estimation (Fan *et al.*, 2012). Although the modified maximum item information method has previously been applied in computerized classification testing to classify each examinee into one of multiple proficiency groups (Sie *et al.*, 2015), the item-selection procedure used in that study was based on IRT models and the Fisher information. In this study, the PWKL information method was used in the item selection process to investigate the effects of incorporating RT information on the accuracy of latent attribute and latent trait estimation and on the control of speededness in CD-CAT.

The modified procedure developed by Fan *et al.* (2012) reduces the amount of time that examinees require to respond to test items; however, it does not guarantee that no examinee exceeds the time limit. Alternatively, a shadow test can be assembled from the item pool to maximize the PWKL information given the current estimate of $\hat{\boldsymbol{\alpha}}$ for the test taker and to ensure simultaneously that the time constraint is satisfied using binary integer programming. Let t_{lim} be the time limit and let x_j be the decision variable ($x_j = 1$ if item j is selected and $x_j = 0$ otherwise); let S_{n-1} be the set of $n-1$ already administered items, let L be the pre-specified test length for the adaptive test and let t_{ij}^* be the actual time spent by examinee i in responding to item j. Using the STA for adaptive testing (van der Linden, 2009a; van der Linden & Xiong, 2013), the model for the assembly of the shadow test can be formulated as

$$\text{maximize} \sum_{j=1}^{J} \text{PWKL}_j(\hat{\alpha}_i) x_j, \tag{21}$$

subject to

$$\sum_{j\in R_n} E(t_j|\hat{\tau}_i)x_j + \sum_{j\in S_{n-1}} t^*_{ij}x_j \leq t_{\text{lim}}, \tag{22}$$

$$\sum_{j\in S_{n-1}} x_j = n-1, \tag{23}$$

$$\sum_{j=1}^{J} x_i = L \tag{24}$$

and

$$x_j \in \{0,1\},\ j = 1,\ldots,J. \tag{25}$$

In addition to permitting the specification of a time constraint, the model is sufficiently flexible to accommodate the specification of other constraints to ensure that an adaptive test respects various practical demands, such as item-exposure control, test-overlap control and content-distribution constraints. The commercial program CPLEX (International Business Machines Corporation, 2015) can be used to generate optimal shadow tests and has been found to be a fast and efficient solver even when hundreds of constraints are imposed (van der Linden, 2009a). Therefore, this efficient solver was adopted in this study for test-assembly optimization in CD-CAT.

The information obtained from RT data can be used not only to prevent examinees from exceeding the time limit but also to improve latent ability estimation. As mentioned above, the level-1 measurement models for item responses and RTs can be flexibly defined by researchers and are not limited to any specific response or RT functions. At a higher level, a joint population distribution can be postulated for the second-order latent trait in the higher-order DINA model and the speed parameter in the lognormal RT model using the hierarchical modelling approach. Because the mastery status for each first-order latent attribute is determined by the second-order latent trait, the speed parameter has indirect effects on the attribute mastery status through the correlation between the θ and τ parameters. Using empirical Bayesian procedures, we obtain the posterior predictive distribution of θ_i conditional on the RTs after $(n-1)$ items have been administered to examinee i as

$$f(\theta_i|\mathbf{t}_{i,n-1}) = \int f(\theta_i|\tau_i)f(\tau_i|\mathbf{t}_{i,n-1})d\tau_i, \tag{26}$$

where the mean and variance can be asymptotically approximated by

$$\mu_{\theta_i|\mathbf{t}_{i,n-1}} = \frac{\sigma_{\theta\tau}\sum_{j=1}^{n-1}\eta_j^2(\delta_j - \ln t_{ij})}{1+\sigma_\tau^2\sum_{j=1}^{n-1}\eta_j^2} \tag{27}$$

and

$$\sigma^2_{\theta_i|\mathbf{t}_{i,n-1}} = 1 - \frac{\sigma^2_{\theta\tau}}{\sigma^2_\tau} + \frac{\sigma^2_{\theta\tau}}{1+\sigma^2_\tau\sum_{j=1}^{n-1}\eta_j^2}, \tag{28}$$

respectively (van der Linden, 2008). Because the conditional distribution of $f(\theta_i|\tau_i)$ is normal, combining it with the normal posterior of τ can result in a normal posterior distribution of θ_i. When the test begins, the prior distribution of θ is assumed to follow a standard normal distribution. As the test continues, the posterior distribution of θ is updated and can be used as the empirical prior distribution of θ to improve the measurement precision. In higher-order CD-CAT with RTs, a joint prior distribution of θ and τ is assumed, and the posterior predictive distribution of the second-order latent trait can be easily derived and implemented. Therefore, in addition to the comparison of different item-selection methods for speededness control, the second principal purpose of this study was to examine, via simulation, the effectiveness of using the RTs as additional information for the estimation of θ under the higher-order DINA model.

As a CD-CAT exam progresses, an examinee might run out of time even if the RTs are considered during the item-selection process. This is because the modified PWKL method (equation [20]) is subject to some uncertainty regarding the expected RTs of the candidate items and does not directly control for test speededness, as in the STA. To conserve resources and to minimize the burden on the respondents, it is desirable to ensure that the percentage of examinees who exceed the given time limit remains within an acceptable range. Let π be the risk (or probability) of a test taker running out of time, and let $T^*_{i,n-1}$ be the time remaining after $(n-1)$ items have been administered to examinee i. The predictive probability that item j will require more than the remaining time conditional on the RTs for the previous $(n-1)$ items can be expressed as

$$P(t_{ij} > T^*_{i,n-1}|\mathbf{t}_{i,n-1}) = \int P(T_{ij} > T^*_{i,n-1}|\tau_i) f(\tau_i|\mathbf{t}_{i,n-1}) d\tau_i, \tag{29}$$

where CD-CAT terminates when the computed predictive probability is greater than π. Sie *et al.* (2015) have found that such a termination procedure can reduce testing time, while simultaneously resulting in higher classification accuracy in computerized classification testing. Therefore, this procedure for terminating testing when the time required for the next item is expected to exceed the remaining time was adopted and coupled with the modified PWKL method to investigate the effectiveness of the newly proposed CD-CAT algorithms under the higher-order DINA model.

Although using RT information to control speededness and improve latent trait estimation precision for examinees has been well documented in the IRT-based CAT literature, few studies have incorporated RT information into CD-CAT and assessed the utility of RT information in RT control and attribute mastery classification. Intuitively, the conclusions drawn from IRT-based CAT studies with the use of RTs are expected to be similar to those obtained from the results of the CD-CAT scenario; however, the way in which factors (e.g., attribute number, test length, association of latent trait with speed parameters, and fixed- or variable-length condition) potentially affect the precisions of $\boldsymbol{\alpha}$ and θ estimation and the effectiveness of RT control remains unclear. This is mainly because IRT-based CAT and CD-CAT are developed with different methodological perspectives and measurement purposes. In addition, when the higher-order DINA model is adopted as the psychometric model in CD-CAT, imprecise latent continuous trait estimation would compromise the inference validity of summative assessment. Because RT data are easily accessible in any computer-based testing, joint hierarchical modelling of responses and RTs helps to exploit such information in person–parameter estimation in CD-CAT. The hierarchical model framework can be constructed based either on the association of the latent trait parameter and the speed parameter or on the association of

the latent attributes and the speed parameter. This means that this hierarchical modelling approach is more flexible than those in the literature. Furthermore, fixed- and variable-length CD-CAT were considered for the use of RT information in item selection, attribute classification and latent trait estimation in this study, which have never been discussed in the CD-CAT literature. All the above reasons constitute the most significant contribution of this article to testing practices and the theoretical field.

6. Simulation design

6.1. Scenario 1: CD-CAT using the higher-order DINA model with the use of RT information

Two simulation studies were designed to evaluate the benefit of utilizing the additional information obtained from an examinee's RTs in the item-selection process in order to control the speededness of the test and to improve the accuracy of latent ability estimation in CD-CAT under the higher-order DINA model. In the first simulation study, a fixed-length CD-CAT framework was used, and several approaches to administering items, both considering and not considering RTs, were compared: the original PWKL method, which does not consider the RTs of the examinees; the speededness-control PWKL method (abbreviated as the SC method), which searches for the item that yields the maximum information per time unit; and the shadow-test method (abbreviated as the ST method). In addition, the latter two approaches use the empirical prior distribution of θ derived from the RT data as prior information for estimating the second-order latent trait in the higher-order DINA model, whereas the first approach does not. The test length was set to 10, 15 or 20 items for the fixed-length CD-CAT study.

In the second simulation study, which focused on variable-length CD-CAT, the measurement precisions for both attribute and latent trait estimation were considered. Following the recommendations of Hsu and Wang (2015), I defined two termination rules: the first rule was defined based on criteria of $P(\hat{\boldsymbol{\alpha}}_{\max}) = 0.80$ and $P_{\theta}(\hat{\boldsymbol{\alpha}}_c) = 0.85$ to represent a lenient termination condition, and the second rule, defined by $P(\hat{\boldsymbol{\alpha}}_{\max}) = 0.90$ and $P_{\theta}(\hat{\boldsymbol{\alpha}}_c) = 0.95$, represented a strict termination condition. For each termination rule, three item-selection methods were implemented: one was the original PWKL method, and the other two were variants on the speededness-control PWKL method. In the first speededness-control PWKL method (abbreviated as the SC-1 method), the empirical prior distribution of θ is used to improve the second-order latent trait estimation, whereas this is not done in the second method (abbreviated as the SC-2 method). For all three methods, the time-limited stopping rule defined in equation (29) was used to terminate testing when the time required for the next item was expected to exceed the remaining time. Because the ST method requires the specification of a fixed-form test to produce a shadow test, this method is not applicable in variable-length CD-CAT and, consequently, was not investigated in this study. The designs for both simulation studies, in terms of the incorporation of RT information during the item-selection process, the latent trait estimation and the stopping criterion, are presented in Table 1.

For both simulation studies, the higher-order DINA model was used as the CDM for the CD-CAT framework, in which an item pool consisting of 300 items measuring five or seven attributes was simulated. The location parameters λ_{0k} were set to -2.00, -1.00, 0, 1.00 and 2.00 for the five-attribute condition and to -2.00, -1.33, -0.67, 0, 0.67, 1.33 and 2.00 for the seven-attribute condition. The discrimination parameters λ_{1k} were sampled from the lognormal (0, 0.0625) distribution, and the generated values were 0.97, 0.90, 1.01,

Table 1. Simulation designs for fixed-length and variable-length CD-CAT studies.

	Process using RT information		
CD-CAT type	Item selection	Latent trait estimation	Termination
Fixed-length			
PWKL method			
SC method	V	V	
ST method	V	V	
Variable-length			
PWKL method			V
SC-1 method	V	V	V
SC-2 method	V		V

1.02 and 0.93 for the five-attribute item bank and 0.98, 1.02, 1.06, 1.03, 0.97, 0.92 and 0.96 for the seven-attribute item bank. The slip and guessing parameters were generated from a uniform distribution ranging between 0.05 and 0.25, and the same generated values were used for both the five- and seven-attribute conditions. These settings for the attribute- and item-parameter distributions were representative of those that are commonly observed in real data analyses, and were consistent with or similar to those used in previous studies (Chen, Xin, Wang, & Chang, 2012; de la Torre & Douglas, 2004; Hsu & Wang, 2015; Hsu *et al.*, 2013; Huang & Wang, 2014; Kaplan *et al.*, 2015; Mao & Xin, 2013; Wang, 2013). The **Q**-matrix was generated with reference to previous studies (Chen, Liu, & Ying, 2015; Chen *et al.*, 2012), in which three basic matrices that specify all possible patterns for items measuring one, two and three attribute(s) were generated and then randomly reordered to constitute the full **Q**-matrix.

For the RT model, the discrimination parameters η_i were generated from the $U(1, 3)$ distribution. Let δ_{ik} be the attribute-level time-intensity parameter for attribute k and item i, and let δ_i be the item-level time-intensity parameter for item i. The parameters δ_{ik} were assumed to be correlated with the location parameters λ_{0k} and to follow a normal distribution with a mean of 4 and a variance of one-third. The correlation between the parameters δ_{ik} and λ_{0k} was set to 0.65. Consequently, the parameters δ_{ik} were randomly sampled from the corresponding conditional distribution given the parameters λ_{0k}, which can be expressed as $N(4 + 0.65\sqrt{1/3}/\lambda_{0k}, 1/3(1 - 0.65^2))$. After all parameters δ_{ik} were generated, I computed the item-level time-intensity parameter as $\delta_i = \sum_{k=1}^{K} q_{ik}\delta_{ik} / \sum_{k=1}^{K} q_{ik}$ for each item. For example, if an item with a $\boldsymbol{q}$-vector of (1,1,0,1,0) had corresponding δ_{ik} values of 2.66, 3.38, 4.02, 3.95 and 5.15 for the five attributes represented by the vector, respectively, then a value of 3.33 was generated for δ_i. Similar simulation designs and generated values for the RT model have been reported in previous studies (Sie *et al.*, 2015; van der Linden, 2009a; van der Linden, Scrams, & Schnipke, 1999).

Two levels of correlation, low and high, between the latent trait parameter θ and the speed parameter τ were defined, using $\rho_{\theta\tau} = 0.40$ and $\rho_{\theta\tau} = 0.80$, respectively (van der Linden, 2008, 2009a). Nine true values of the latent trait parameter were simulated, namely, $\theta = -2.0, -1.5, \ldots, 2.0$, with 500 examinees (replications) for each. The θ parameter was assumed to follow a standard normal distribution, and the τ parameter was assumed to be normally distributed with a mean of zero and a variance of 0.35 (van der Linden, 2009a; van der Linden & Xiong, 2013). Given the true latent ability, a separate

value of τ was sampled for each replication under the assumption of a bivariate normal distribution with a zero mean vector and a variance-covariance matrix of $\begin{pmatrix} 1 & \rho_{\theta\tau}\sqrt{.35} \\ \rho_{\theta\tau}\sqrt{.35} & .35 \end{pmatrix}$. Specifically, given each of nine θ points, the τ parameters were generated by sampling a separate value from its conditional distribution of $N(\rho_{\theta\tau}\sigma_\tau\theta, \sigma_\tau^2 - \rho_{\theta\tau}^2\sigma_\tau^2)$. Values similar to the τ parameter generation in this study can be found in previous studies (Sie *et al.*, 2015; van der Linden, 2008).

The setting of the time limit for a CD-CAT exam is an important and practical issue. In the fixed-length CD-CAT scenario, the time limits were set to 1,160, 1,740 and 2,320 s for the 10-, 15-, and 20-item test lengths, respectively, based on the 15-item Armed Services Vocational Aptitude Battery adaptive test, for which the actual time limit is 1,740 s and the average expected RT per item is 116 s, based on similar time limits used in previous studies (van der Linden, 2009a; van der Linden *et al.*, 1999). The setting of the time limit for the second simulation study was more complicated than that for the first simulation study because the test length was variable for each examinee. Therefore, a preliminary simulation was conducted to determine how the time limit should be set for the variable-length CD-CAT scenario; in this preliminary simulation, $\rho_{\theta\tau} = 0.60$ was set for the generation of the data, but the PWKL method (without RTs) was used for item selection, and the RT priors were not used for latent trait estimation. These conditions represent a moderate correlation between the θ and τ parameters. The average RT spent by examinees on the administered items was computed for use as the time limit, and four time limits were determined, corresponding to the four different combinations of the two numbers of attributes of interest (i.e., five and seven attributes) and the two stopping criteria (i.e., the lenient and strict termination rules). Consequently, when five attributes were measured, the time limits were set to 582 and 674 s for the lenient and strict termination rules, respectively, and when seven attributes were measured, the time limits were set to 757 and 920 s, respectively, for these two termination rules.

To evaluate the CD-CAT quality, the correct classification rates (CCRs) for individual attributes (as the mean across all attributes) and for the entire latent class profile were calculated by averaging over the 500 replications at each level of θ. For the estimators of the latent trait parameter θ and the speed parameter τ, the RMSEs were computed to assess the efficacy of parameter recovery. The percentage of examinees who exceeded the pre-specified time limit was also recorded to assess the efficacy of the proposed method in controlling the speededness of the CD-CAT exam.

For the two simulation studies, the following results were expected. In the fixed-length CD-CAT scenario, both the SC and ST methods were expected to provide satisfactory attribute mastery recovery comparable to that of the PWKL method, with the advantages of simultaneously improving latent trait estimation and decreasing the risk of running out of time. The superiority of both the SC and ST methods over the PWKL method was expected to become more prominent as the correlation between the θ and τ parameters increased. In the variable-length CD-CAT scenario, the SC-1 and SC-2 methods, in which RT information is used during item selection, were expected to yield results similar to those of the PWKL method in terms of attribute estimation, but the SC-1 method was expected to perform better than the SC-2 method in terms of the recovery of the latent trait parameters because the empirical prior distribution of θ and τ was considered only in the SC-1 method. Setting a more stringent time limit was expected to reduce the efficacy of the new methods in removing the risk of running out of time for the examinees.

6.2. Scenario 2: CD-CAT using the DINA model with the use of RT information

Rather than using the higher-order CDM approach, the association between α and τ can be directly modelled (see equation [9]), and the information borrowed from the τ parameter can be used in the prior distribution when classifying examinees into the mastery/non-mastery status in CD-CAT. The direct effect of the speed parameter on the α parameter was assessed by manipulating the third simulation study. A fixed-length CD-CAT exam of 10 items in a 300-item pool measuring five attributes was administered to 2,000 examinees, and the SC method was selected as the item-selection method. The attribute mastery status of examinees for the five attributes (i.e., the prior probability) was regressed on the τ parameter generated from a standard normal distribution, where two regression weights of 0.8 and 0.4 were varied (i.e., the five attributes had regression weights of either 0.8 or 0.4) and the intercept was fixed to zero for the five attributes. Attribute estimations with and without RT information were compared in terms of attribute classification accuracy. Note that regardless of whether attribute estimation depended on RT information, the speed parameters should be estimated using the lognormal RT model because the same speed-control PWKL method was used for item selection in both comparison methods.

In addition, two types of item-parameter quality were manipulated: the high-quality parameter setting had slip and guessing parameter values between 0.05 and 0.25 (the same as in Scenario 1), whereas the low-quality parameter setting had values between 0.25 and 0.45. The **Q**-matrix and the RT model parameters were set to be identical to the simulations in Scenario 1. Each condition was replicated 25 times to control for possible sampling variation.

When examinees are administered a CD-CAT exam with the use of RT information, their speed parameter estimates are updated iteratively, and the probability of each attribute mastery pattern can be estimated (see equation [11]). Therefore, the uniform prior used for $P(\boldsymbol{\alpha}_c)$ in equation (13) can be replaced by the estimated probability obtained from the prediction function. I expected that, at the early stage of CD-CAT, the individualized prior probabilities of attribute patterns would improve the quality of attribute classification whereas, as the CD-CAT proceeded, the effects of the prior information would decline and the difference between the uniform and RT-based priors in attribute mastery recovery would become trivial.

7. Results

7.1. Scenario 1: CD-CAT simulation using the higher-order DINA model with the use of RT information

7.1.1. Fixed-length CD-CAT

Figure 1 compares the different CD-CAT algorithms in terms of the CCR for individual attributes as a function of the θ level in the five-attribute condition for the various combinations of the values of ability–speed correlation and test length. The three methods exhibited similar and satisfactory individual attribute-recovery performances, although a slight difference among the three methods in terms of their average CCRs was observed for the shortest test length of 10 items. As the test length increased, the attribute mastery statuses of examinees could be more accurately classified, and the differences among the three methods diminished. There were no systematic differences between the high and low ability–speed correlations. The same conclusions held for the recovery of the entire

attribute profile, as shown in Figure 2, although the CCRs for the entire attribute profile were lower.

Regarding the accuracy of latent continuous trait estimation, the plots of the RMSEs of the θ estimates that are presented in Figure 3 show that both the SC and ST methods, which use RT information, outperformed the PWKL method, which does not use RT information. The plots also show that as the ability–speed correlation increased and as the abilities of the test takers approached the extreme latent trait levels, the superiority of the SC and ST methods over the PWKL method became more prominent. The test length appeared to have little impact on the latent trait estimation. These results provide a potent justification for using the examinees' RTs on test items as supplementary information because they show that doing so can substantially improve the latent trait estimation in a higher-order CD-CAT exam.

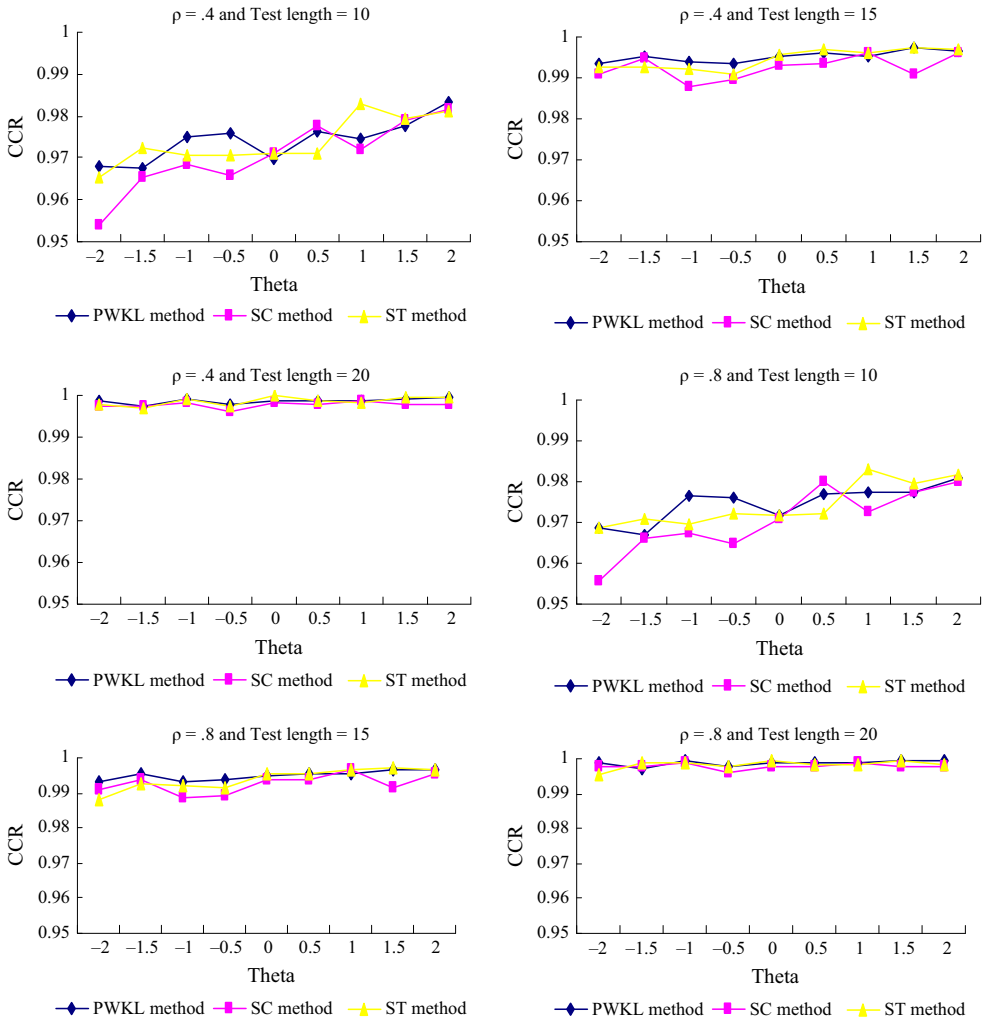

Figure 1. CCRs for individual attributes for the item pool measuring five attributes in the fixed-length CD-CAT scenario. *Note:* The average CCRs across all attributes were computed to evaluate the recovery of individual attribute mastery. [Colour figure can be viewed at wileyonlinelibrary.com]

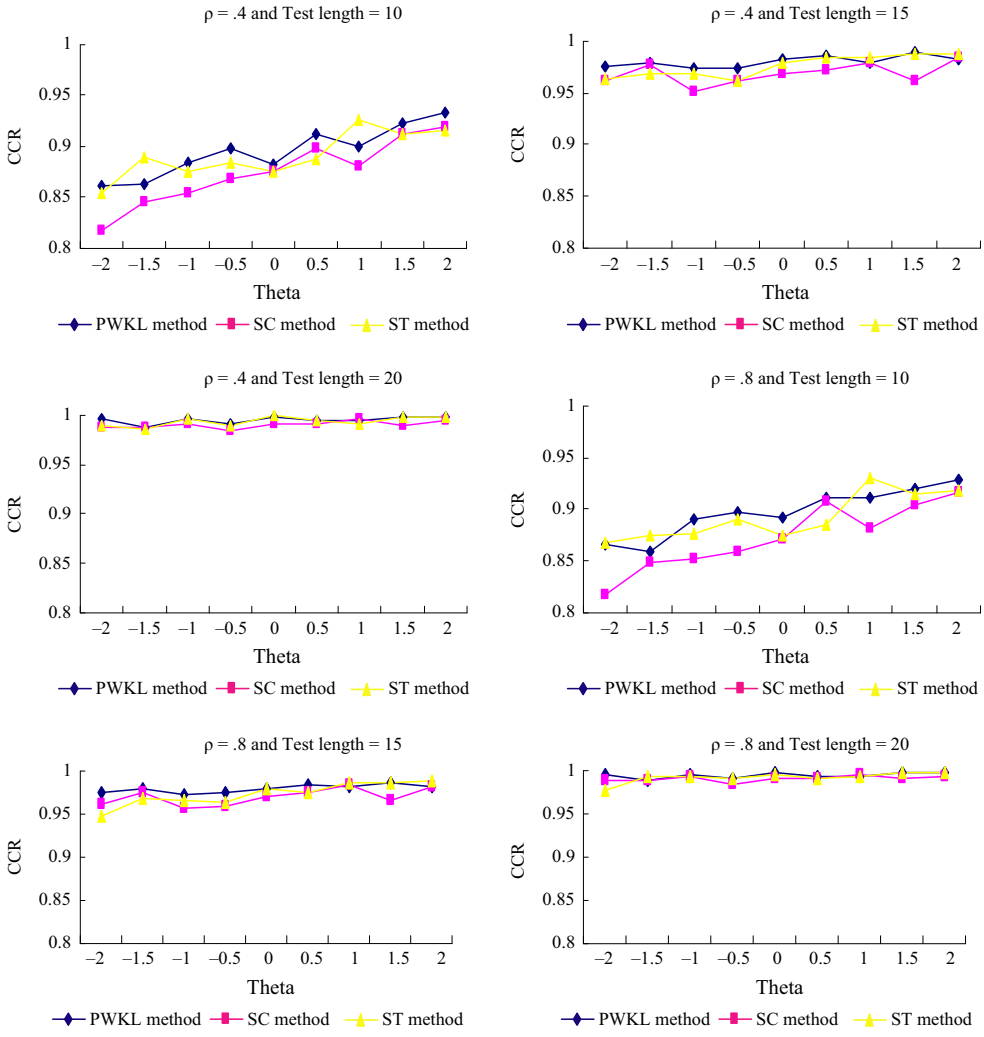

Figure 2. CCRs for the entire attribute profile for the item pool measuring five attributes in the fixed-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 4 presents the RMSEs of the τ estimates, which show that the speed parameter can be estimated more accurately by using either the SC or ST method rather than by using the PWKL method, regardless of the ability–speed correlation or the test length, and that the RMSE is independent of the trait level of the test takers. However, the test length did exert an effect on the accuracy of the speed parameter estimation, with a longer test length yielding smaller RMSE values.

With respect to the effectiveness of speededness control, Figure 5 plots the percentage of examinees who exceeded the pre-specified time limit as a function of the θ level for all three methods. Because the ability–speed correlation was set to 0.40 or 0.80 (positive values), examinees with lower θ values were expected to have a higher likelihood of exceeding the time limit. Obviously, both the SC and ST methods could ensure that the risk of running out of time remained lower (in fact, close to zero), whereas the PWKL method resulted in higher percentages of examinees exceeding the time limit

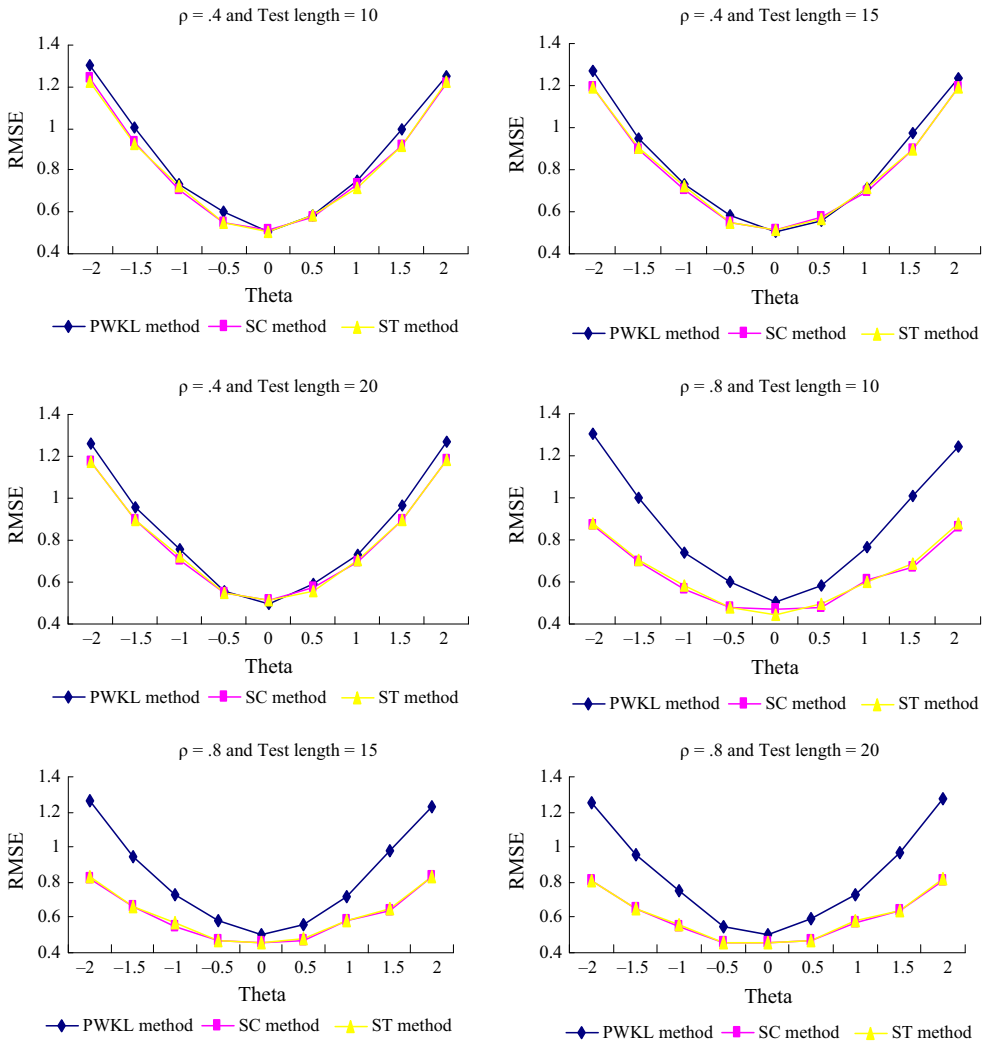

Figure 3. RMSEs of latent trait parameter estimates for the item pool measuring five attributes in the fixed-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

in all cases. Furthermore, the difference in percentage between the SC and ST methods and the PWKL method became larger when the ability–speed correlation increased. No systematic patterns with respect to test length could be identified because different time limits were specified for different test lengths.

Because of spatial limitations and because the results for the seven-attribute condition were comparable to those for the five-attribute condition, detailed results for the former are not reported here. However, some differences between the five- and seven-attribute conditions were noted, as follows. The quality of attribute estimation for the item pool measuring seven attributes was inferior to that for the item pool measuring five attributes, because the seven-attribute condition required more test items to attain the same level of classification accuracy compared with the five-attribute condition. Conversely, because the accuracy of latent trait estimation strongly depends on the number of attributes, the

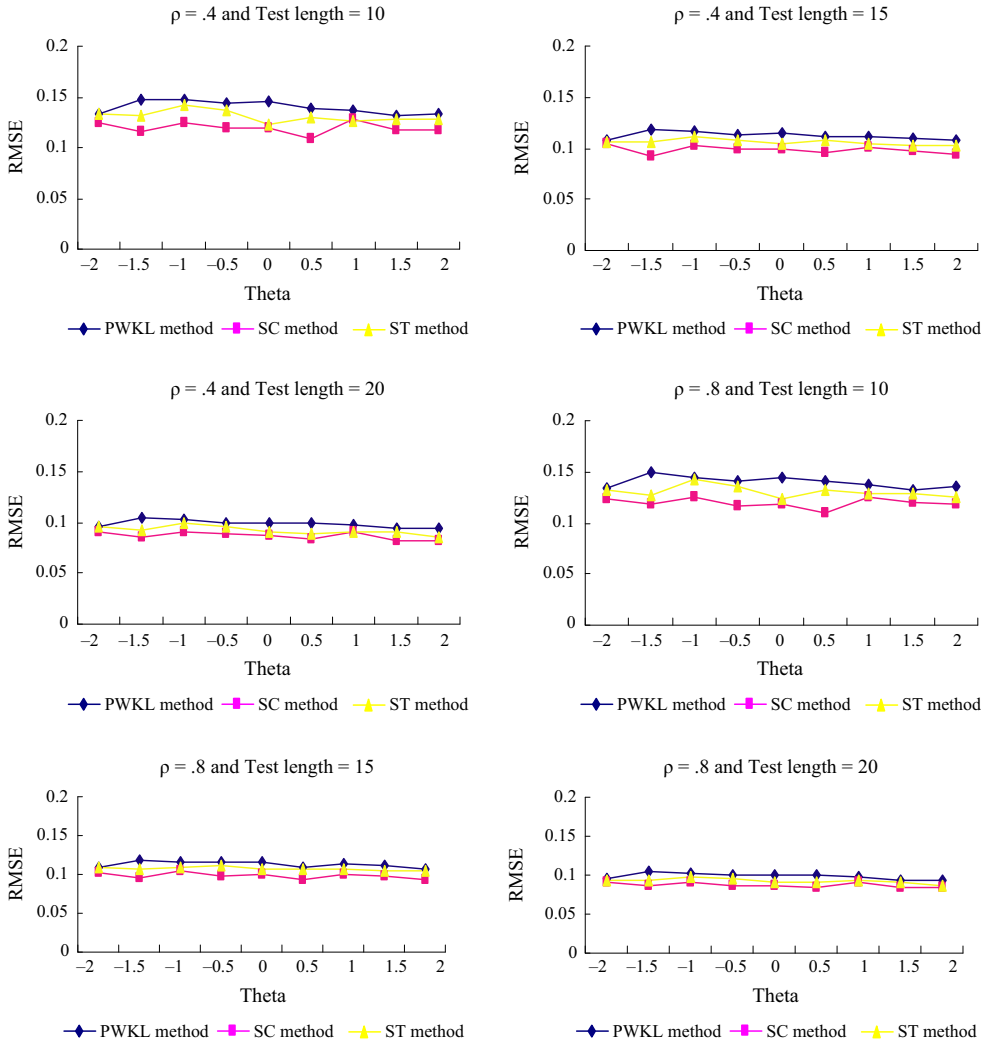

Figure 4. RMSEs of speed parameter estimates for the item pool measuring five attributes in the fixed-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

RMSEs of the θ estimates were found to be slightly lower in the seven-attribute condition than in the five-attribute condition. All results are available from the author upon request.

7.1.2. Variable-length CD-CAT

When the variable-length CD-CAT exam was administered to examinees, lenient (i.e., $P(\hat{\boldsymbol{\alpha}}_{\max}) = 0.80$ and $P_\theta(\hat{\boldsymbol{\alpha}}_c) = 0.85$) and strict (i.e., $P(\hat{\boldsymbol{\alpha}}_{\max}) = 0.90$ and $P_\theta(\hat{\boldsymbol{\alpha}}_c) = 0.95$) termination rules together with the time-limited stopping rule were applied to terminate the test for each examinee. First, the condition of an item pool measuring five attributes was investigated. Figure 6 compares the three methods in terms of individual attribute classification accuracy and shows that for each θ level, the SC-1 method was comparable to the SC-2 method and that the PWKL method yielded the lowest CCRs. The correlation

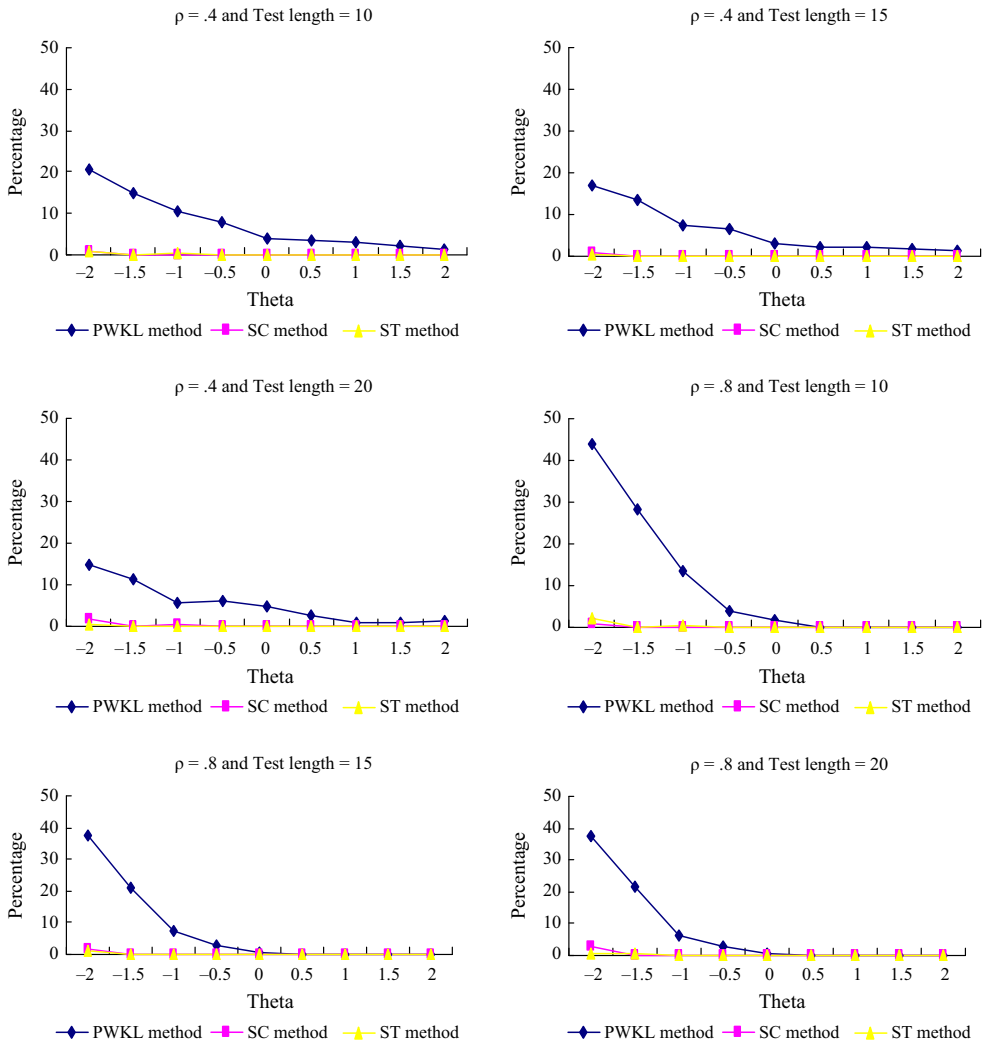

Figure 5. Percentages of examinees who exceeded the pre-specified time limit in the fixed-length CD-CAT scenario when five attributes were measured. [Colour figure can be viewed at wileyonlinelibrary.com]

between the θ and τ parameters had little impact on individual attribute recovery, as in the first simulation study. As expected, the strict termination rule resulted in better attribute recovery than the lenient termination rule, but at the price of a longer test length (as shown in Figure 7). In addition, the SC-1 method appeared to result in the administration of more test items than the other two methods because both the item-selection and ability-estimation procedures considered the RT information. The PWKL method resulted in the shortest test length because time was more likely to run out. Figure 8 shows the CCRs for the entire attribute profile, for which the main conclusions drawn based on the individual attribute-recovery performance still apply.

The RMSEs of the θ estimates, as shown in Figure 9, exhibited patterns similar to those observed in the first simulation study: the SC-1 method yielded the most accurate ability estimates, followed by the SC-2 method and the PWKL method, and the differences were

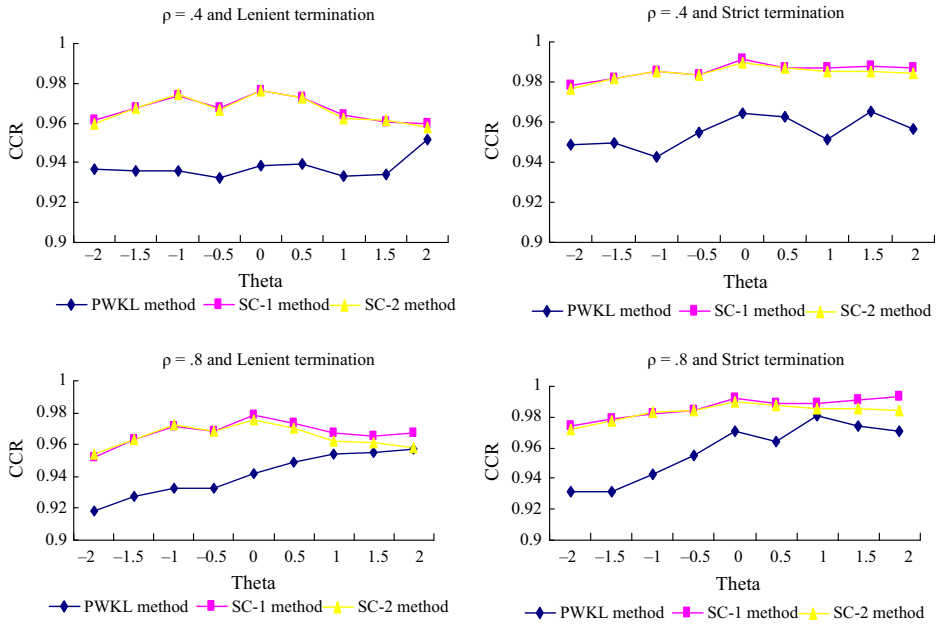

Figure 6. CCRs for individual attributes for the item pool measuring five attributes in the variable-length CD-CAT scenario. *Note:* The average CCRs across all attributes were computed to evaluate the recovery of individual attribute mastery. [Colour figure can be viewed at wileyonlinelibrary.com]

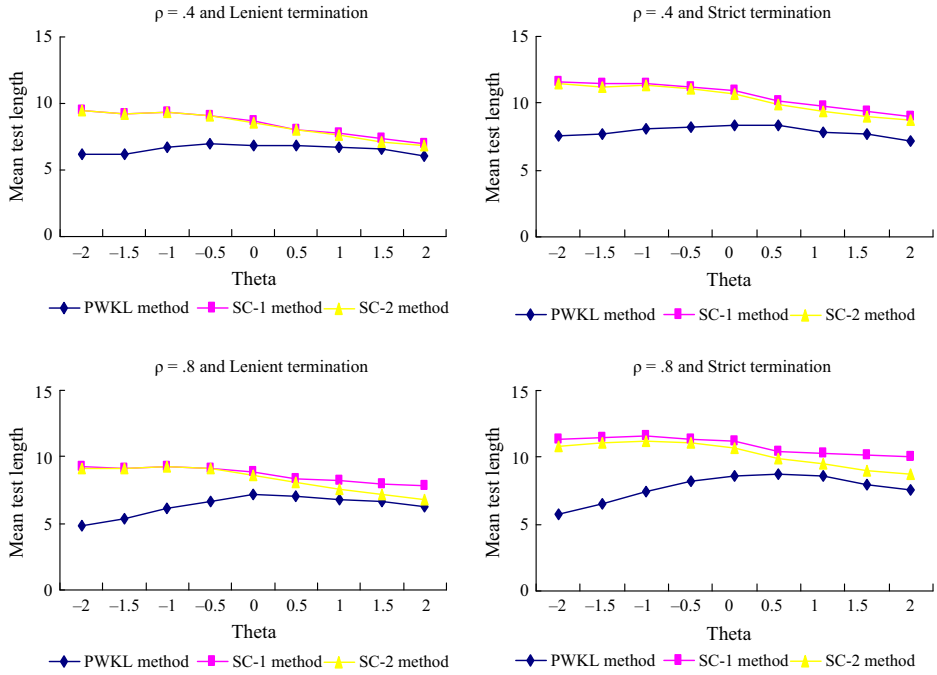

Figure 7. Mean test length administered to examinees for the item pool measuring five attributes in the variable-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

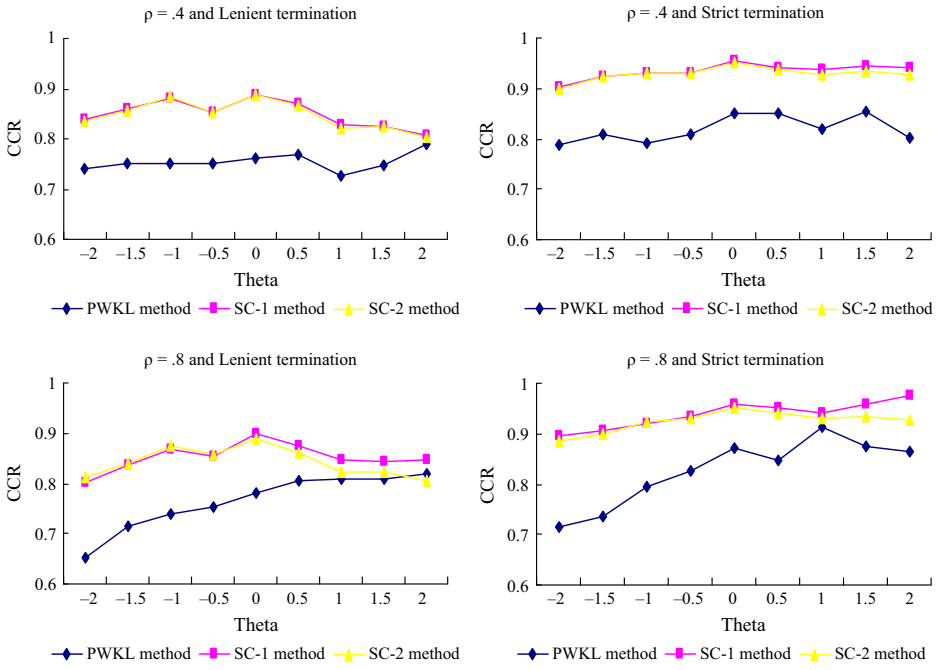

Figure 8. CCRs for the entire attribute profile for the item pool measuring five attributes in the variable-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

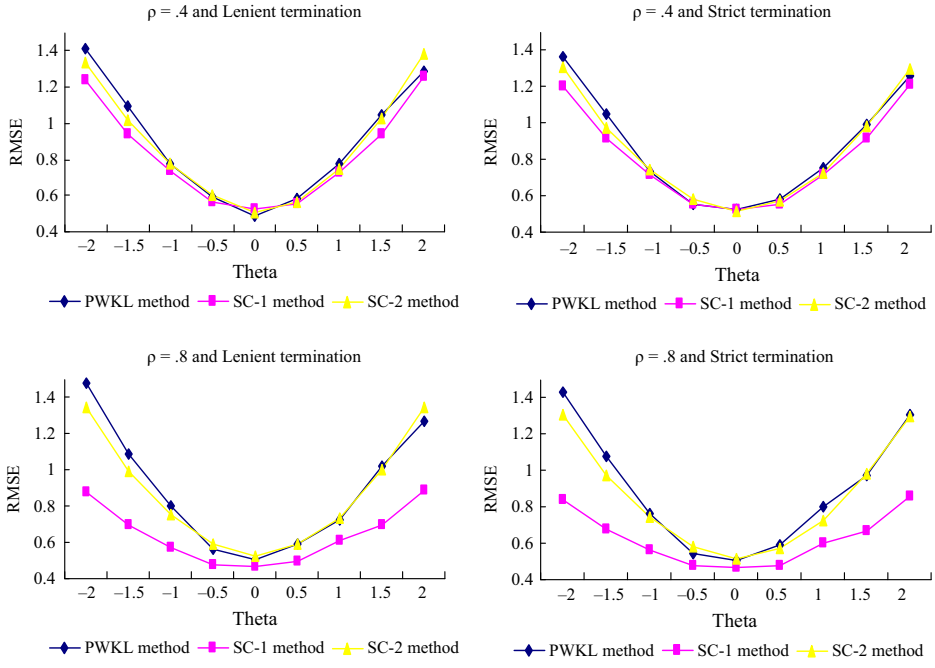

Figure 9. RMSEs of latent trait parameter estimates for the item pool measuring five attributes in the variable-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

more prominent for more extreme trait levels. The SC-1 method outperformed the other two methods because it uses the posterior predictive distribution of θ conditional on the RTs as an empirical prior to improve the second-order latent trait estimation. As the correlation between the θ and τ parameters increased, the superiority of the SC-1 method became more evident. The speed parameter values were more effectively recovered in both the SC-1 and SC-2 methods than in the PWKL method, as shown in Figure 10, and the ability–speed correlation and the termination rule had little impact on the speed parameter estimation.

Figure 11 shows the percentage of examinees who exceeded the time limit as a function of θ for each of the three methods for different ability–speed correlations and termination rules. Note that the percentages were calculated based on not only the number of examinees who exceeded the time limit but also the number of examinees who were expected to have insufficient time to respond to the next item, according to the time-limited stopping rule (i.e., equation [29]). Although the percentages shown in the plots are nearly all increased when compared with those observed in the fixed-length CD-CAT scenario, the SC-1 and SC-2 methods, which consider the RTs during the item-selection process, resulted in a lower risk of running out of time than the PWKL method. Because the time limit was determined based on $\rho_{\theta\tau} = 0.60$, the condition of $\rho_{\theta\tau} = 0.80$ was subject to a relatively stringent time limit and a higher risk of running out of time compared with the condition of $\rho_{\theta\tau} = 0.40$. When the strict termination rule was used, it was found that the examinees were more likely to exceed the time limit than in the case of the lenient termination rule because the strict termination rule required more test items to achieve the minimum accepted measurement precision. Patterns similar to those seen for

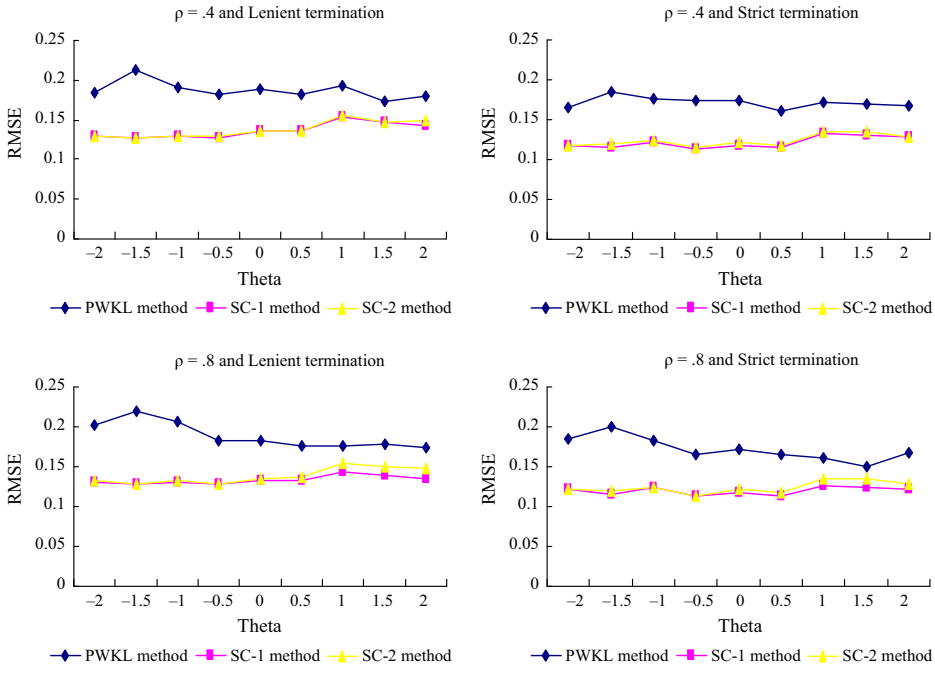

Figure 10. RMSEs of speed parameter estimates for the item pool measuring five attributes in the variable-length CD-CAT scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

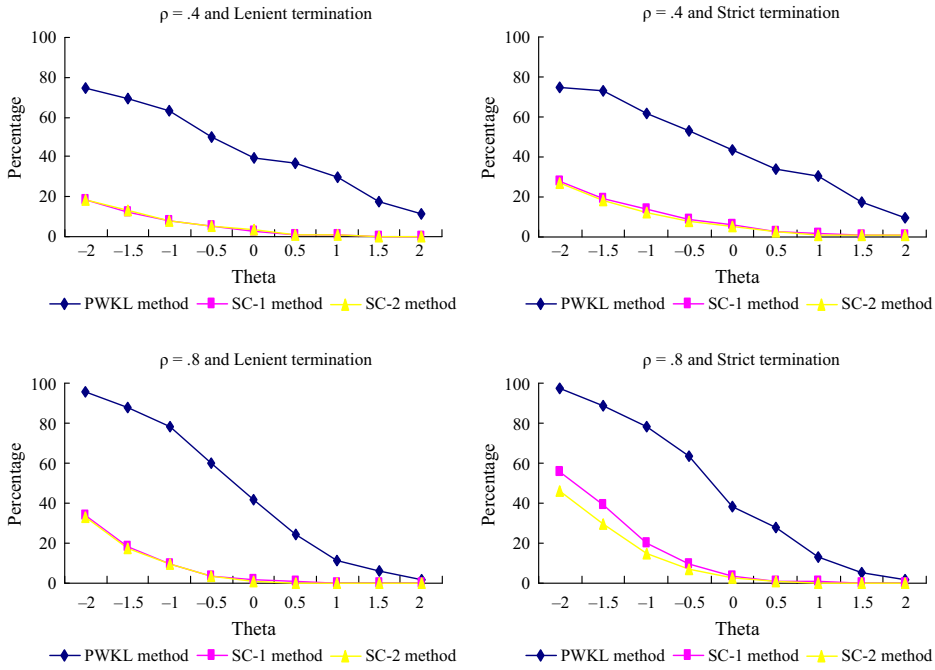

Figure 11. Percentages of examinees who exceeded the pre-specified time limit in the variable-length CD-CAT scenario when five attributes were measured. [Colour figure can be viewed at wileyonlinelibrary.com]

the five-attribute item pool were observed when the seven-attribute item pool was used; the differences between the five- and seven-attribute conditions were the same as in the first simulation study. Given this overlap, detailed results are again not shown, but they are available upon request.

7.2. Scenario 2: CD-CAT simulation using the DINA model with the use of RT information

In this section, the DINA model with RT information was used to simulate the CD-CAT exam, and the relationship between the speed parameter and latent attributes was directly considered rather than addressed via the higher-order modelling approach. Figure 12 shows the comparison between the two CD-CAT algorithms (i.e., the two algorithms differed in whether the estimated speed parameter was used as the predictor for the priors of the latent classes) in the attribute classification accuracy of individual attributes and attribute profiles for the combinations of different regression weights and item qualities. Note that both methods used the modified PWKL information, which considered RTs of examinees for item selection. The results of the comparison indicate how the speed parameter affects attribute estimation, regardless of the second-order continuous latent trait estimation.

When the first item was administered, the same uniform priors for the latent classes were used for each examinee; thus, there was no difference between the two methods in classification accuracy. As the test proceeded, because the speed parameter estimate was

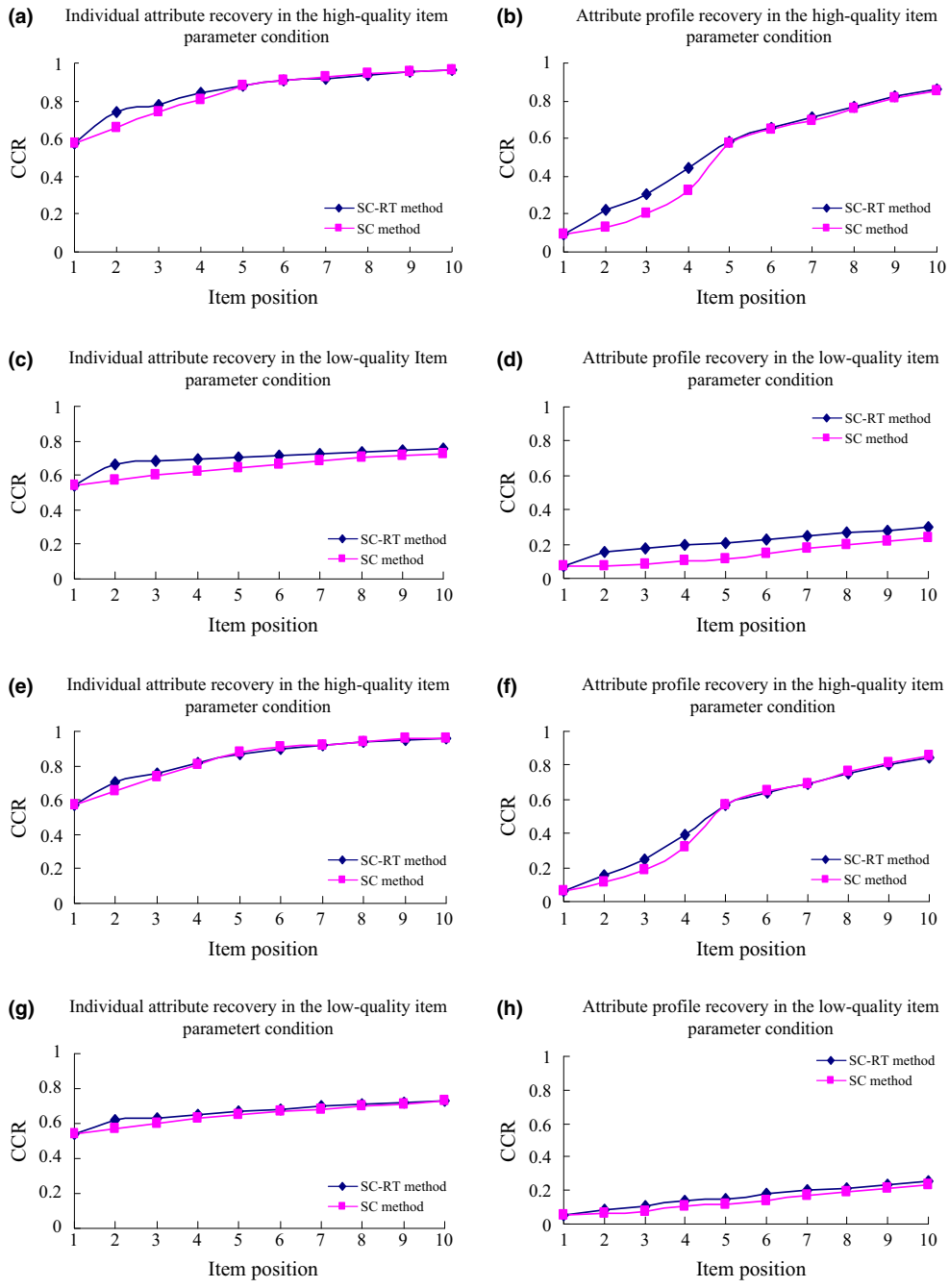

Figure 12. CCRs as a function of item position for the regression weight of 0.8 (panels a–d) and 0.4 (panels e–h) in the fixed-length CD-CAT scenario. Note. Test length = 10; number of attributes = 5; the SC-RT method incorporated RT information into the priors for the attribute estimation, but the SC method did not; both methods used the modified PWKL information. [Colour figure can be viewed at wileyonlinelibrary.com]

updated for each examinee, the individual-specific priors of latent classes were generated and improved the attribute estimation, but a ceiling effect appeared when the test length approached 10 items. When the regression weight was set to 0.8, as shown in Figures 12(a)–(d), the use of RT information to predict the attribute mastery status had higher CCRs than the method without the predicted priors, and the difference was substantial when the low-quality item pool was used. The ceiling effect was observed in the high-quality item pool because the difference in the attribute recovery between the two methods became trivial after the administration of five items. However, in the low-quality item pool, the attribute mastery priors with the use of RT information continued to maintain superiority over the uniform attribute mastery priors in classification accuracy until the CD-CAT exam was terminated. When the regression weight was set to 0.4, as shown in Figures 12(e)–(h), the difference between the two methods with regard to the CCRs decreased, although the major conclusions from the high-regression-weight condition still applied to the low-regression-weight condition.

To understand the difference in attribute classification accuracy between higher-order DINA modelling (using the association of θ and τ as collateral information) and single-order DINA modelling (using the association of $\boldsymbol{\alpha}$ and τ as collateral information) during CD-CAT administration, I compared the results obtained from the first and third simulations, in which the same conditions manipulated both simulations (i.e., test length = 10 and attribute number = 5); the use of the high-quality parameter, the use of the SC method as the item-selection method and the use of RT information to form the prior were considered to make the different simulation designs comparable. Note that the comparison between the two simulations should be interpreted with caution because the first simulation was conducted conditional on θ values, whereas the third simulation study was not, and the two levels of correlation used in the first simulation (i.e., correlation of $\rho_{\theta\tau} = 0.40$ or $\rho_{\theta\tau} = 0.80$) did not exactly correspond to the two levels of prediction power used in the third simulation (i.e., regression weight of $\gamma_1 = 0.40$ or $\gamma_1 = 0.80$).

As shown in Table 2, the level of correlation between θ and τ had little impact on the attribute classification accuracy of both individual attribute recovery and entire profile recovery; this occurs because the attribute mastery status governed by θ and the discrimination parameters with respect to θ (i.e., factor loading) were generated around unity. However, as regression weight increased, the recoveries of the individual

Table 2. Summary results for correct classification rates in CD-CAT under the higher-order DINA model and single-order DINA model

		Correlation/regression weight	
Model	Correct classification rate	Low	High
Higher-order DINA	Individual attribute	0.971	0.971
	Entire profile	0.874	0.875
Single-order DINA	Individual attribute	0.961	0.963
	Entire profile	0.848	0.857

Notes. The SC method was the item-selection method; RT information was used for attribution estimation; test length = 10; attribute number = 5; the high-quality parameter was used; the average correct classification rates across all attributes were computed to evaluate the recovery of individual attribute mastery.

attributes and the entire attribute profile slightly improved. The trivial improvement in the CCRs of attributes when the regression weight increased from 0.40 to 0.80 in CD-CAT under the single-order DINA model was not surprising because the corresponding mean Nagelkerke R^2 for $\gamma_1 = 0.40$ and $\gamma_1 = 0.80$ was 0.045 and 0.157, respectively, across the five attributes. Therefore, when the regression weights are considerably increased to improve predictive power, we can expect that the attribute classification accuracy would be substantially increased because the direct association of $\boldsymbol{\alpha}$ with τ is considered.

8. Conclusions

When examinees are administered a test by a computer, their RTs can be routinely recorded by custom-made programs and can provide useful information for multiple purposes (Fan *et al.*, 2012; Sie *et al.*, 2015; van der Linden, 2008, 2009a; van der Linden & Xiong, 2013). The main purpose of this study was to develop a new set of CD-CAT algorithms to improve item-selection strategies, to reduce the test duration in order to avoid the risk of running out of time, and to enhance the accuracy of estimation for discrete attributes and continuous traits. Simultaneous calibration of the parameters of the CDM and the RT model in the item pool enables the incorporation of RT information into a CD-CAT exam. The higher-order DINA model was used in this study because the DINA model is the most frequently used CDM in the literature, it is simple to interpret, and its higher-order structure can facilitate both formative and summative assessment.

The first simulation study focused on a fixed-length CD-CAT exam, and the results revealed that both the SC and ST methods, which considered RT information during item selection and ability estimation, outperformed the PWKL method in estimating the latent trait and speed parameters while simultaneously maintaining a low risk of running out of time for examinees. The incorporation of the RT information in the SC and ST methods did not compromise the quality of attribute estimation. In addition, for both the SC and ST methods, as the correlation between the ability and speed parameters increased, the latent continuous trait recovery improved. This study contributes to research on latent trait estimation in higher-order CDMs by presenting a new set of CD-CAT algorithms that improve upon the dissatisfactory quality of latent trait estimation achieved in previous studies (de la Torre & Douglas, 2004; Hsu & Wang, 2015). Compared with the SC method, the ST method offers greater flexibility because other practical constraints, such as item-exposure control, can easily be considered in the assembly of shadow tests.

Conclusions similar to those of the first simulation study were also drawn from the second simulation study, which addressed variable-length CD-CAT. In addition to the test termination criterion that is required in variable-length CD-CAT, in order to improve testing efficiency and to reduce the burden on the respondents, a time-limited stopping rule was imposed as an additional stopping criterion. The results show that both the SC-1 and SC-2 methods performed better than the PWKL method in terms of all dependent variables. The SC-1 method with strict termination is highly recommended because it enables satisfactory attribute estimation and uses an empirical θ prior distribution derived from the joint distribution of the latent ability and speed parameters to improve the accuracy of latent trait estimation. Note that although, in this study, the DINA model and the PWKL method were used as the psychometric model and the item-selection method, respectively, very similar adjustments could be applied for other CDMs (de la Torre, 2011)

and other item-selection strategies (Kaplan *et al.*, 2015). In this case, very similar patterns would be expected.

The third simulation study demonstrated the efficiency of RT information in attribute estimation using the association of the speed parameter with the latent attributes. In contrast to the simulations in Scenario 1, the collateral information was obtained in the shape of the conditional prior distributions for the latent classes given the speed parameter estimate. As expected, the priors based on RT information performed better in the attribute recovery than the uniform priors. However, as the test length increased, the item quality improved and the regression weight decreased, the superiority of the priors based on RT information over the uniform priors diminished. Occasionally, the use of a short test length is preferable to a long test length in cognitive diagnostic tests because the former can prevent unnecessary loss of lecture time (Wang, 2013). Therefore, the incorporation of RT information into the priors in the attribute estimation can be justified by its estimation utility.

A higher-order latent trait was used to model the joint distribution of the attributes in a relatively simple model to account for the relationship between a general aptitude and several specific attributes. This higher-order modelling framework in CDMs offers several advantages. First, the complexity of the saturated CDMs can be substantially simplified in the higher-order CDMs (i.e., $2^K - 1$ parameters can be reduced to $2K$ parameters to account for the relationships among the latent classes). Second, it is beneficial for any educational assessments with diagnostic functions because it helps students to identify their strengths and weaknesses in a set of attributes, it guides instructors in improving their teaching and it provides overall performance evaluations according to some standard or benchmark. The higher-order CDMs can meet these demands because both formative and summative assessments are available. Furthermore, the logit link function used in the higher-order CDMs makes parameter estimation simple (de la Torre & Douglas, 2004; Hsu & Wang, 2015). However, the higher-order CDM approach should be implemented with caution. As suggested by de la Torre and Douglas (2004, p. 338) in their original study, an expert option is important not only to construct the **Q**-matrix but also to decide which discriminating power should be estimated for each attribute. Therefore, it might be the case that some of the attitudes are not correlated highly enough to produce higher discrimination parameters and that there are other nuisance latent variables to interfere with the parameter estimation. Improving the relationship between the attributes and latent continuous trait by selecting appropriate attributes based on cognitive theory is an exercise and a critical responsibility for test developers.

In this study, the speed parameter values were estimated based only on the RTs for previously administered items, regardless of the correctness of the responses to those items. The use of the item responses as an additional source of information for speed estimation to predict the RTs for the remaining items in the pool has been proposed in the context of fixed-length IRT-based CAT (van der Linden, 2009a). However, the specifics of how such a complicated estimation algorithm would function for fixed- and variable-length CD-CAT exams require further investigation. As shown in the second simulation study, the imposition of a stringent time limit reduces the efficiency of the proposed methods, and thus the setting of a reasonable time limit is crucial. An alternative, modified approach is to identify a reference test form with a proven level of speededness and then to attempt to produce an adaptive test with the same speededness as that of the reference form. By doing so, researchers can avoid the problem of specifying the time limit and can prevent any additional RT parameter estimation from being required during test

administration (van der Linden & Xiong, 2013). Determining how to apply this alternative approach in CD-CAT would be an interesting and important topic for future study.

Regarding one anonymous referee's comment on the variable-length CD-CAT scenario, the use of the minimum-precision rule for the second-order latent trait estimate ($\hat{\theta}$) to terminate CD-CAT is not justifiable. I decided to keep the termination rule in this study for several reasons. First, CD-CAT will not stop until the minimum-precision termination criteria for both $\hat{\theta}$ and $\hat{\boldsymbol{\alpha}}$ are satisfied. Therefore, the latent continuous and binary parameters of examinees are expected to be estimated more precisely with the use of both determination rules than with a condition that only implements a single rule to terminate CD-CAT. The same results found in the study of Hsu and Wang (2015) were found in our preliminary study to conduct data for a simulation study that only used the maximum posterior probability to terminate CD-CAT, in which the measurement precision for $\hat{\theta}$ and $\hat{\boldsymbol{\alpha}}$ decreased compared to that of a counterpart that uses both termination rules at the cost of little increase in test length. In addition, previous studies have used information simultaneously about both θ and $\boldsymbol{\alpha}$ to proceed with CAT under higher-order CDMs (McGlohen & Chang, 2008; Wang *et al.*, 2012). Second, as described and highlighted above, the second-order latent trait can be used to assess the overall performance of examinees, and the precision of $\hat{\theta}$ should be seriously considered. In doing so, any inferences for examinees' overall performance will not be compromised. Third, because a small to moderate number of attributes are often used in CDM literature, we cannot expect the θ estimate to be very precise by arbitrarily specifying a stringent minimum-precision criterion for $\hat{\theta}$. Instead, the use of the target latent class to determine the minimum-precision termination rule for $\hat{\theta}$ can help practitioners and researchers specify a reasonable threshold to stop CD-CAT. Finally, although the precision of θ estimation may be limited by the small number of latent attributes, imprecise θ estimation can be compensated for by using RT data as collateral information to overcome the inherent limitation of a small number of attributes.

For simplicity, I have assumed a lack of calibration errors on the parameter estimates for the RT model in this simulation design and I have treated the parameter estimates as if they were the true population parameters. However, because the predictions of the examinees' RTs depend on their speed parameter estimates, which are obtained based on the parameter estimates for the RT model, the effects of calibration errors on attribute and ability estimation and on the effectiveness of speededness control deserve further exploration. Finally, the **Q**-matrix was assumed to be correctly constructed by domain experts; however, the **Q**-matrix should be empirically validated (de la Torre & Chiu, 2016). The influence of misspecification of the **Q**-matrix on the outcomes of CD-CAT methods that incorporate RT information would also be an interesting topic for future research.

Acknowledgements

This study was supported by the Ministry of Science and Technology, Taiwan (Grant No. 106-2628-H-845-001-MY2).

References

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied Psychological Measurement*, *39*, 5–15. https://doi.org/10.1177/0146621613513065

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77, 201–222. https://doi.org/10.1007/s11336-012-9255-7

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632. https://doi.org/10.1007/s11336-009-9123-2

Dai, B., Ding, S., & Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, *19*, 1465–1483. https://doi.org/10.3150/12-bejsp10

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353. https://doi.org/10.1007/bf02295640

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285–291. https://doi.org/10.1111/j.1745-3984.2007.00039.x

Entink, R. H. K., Fox, J. P., & van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48. https://doi.org/10.1007/s11336-008-9075-y

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

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*, 655–670. https://doi.org/10.3102/1076998611422912

Hsu, C.-L., & Wang, W.-C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement*, *52*, 125–143. https://doi.org/10.1111/jedm.12069

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*, 563–582. https://doi.org/10.1177/0146621613488642

Huang, H.-Y., & Wang, W.-C. (2014). The random-effect DINA model. *Journal of Educational Measurement*, *51*, 75–97. https://doi.org/10.1111/jedm.12035

International Business Machines Corporation. (2015). *IBM Ilog CPLEX optimization studio, version 12.6.2 [software program and manual]*. Armonk, NY: International Business Machines Corporation.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*, 167–188. https://doi.org/10.1177/0146621614554650

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, *30*, 152–172. https://doi.org/10.1007/s00357-013-9128-5

Mao, X., & Xin, T. (2013). The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Applied Psychological Measurement*, *37*, 482–496. https://doi.org/10.1177/0146621613486015

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808–821. https://doi.org/10.3758/brm.40.3.808

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*, 197–219. https://doi.org/10.1111/bmsp.12042

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141. https://doi.org/10.1007/s11336-015-9476-7

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement*, *39*, 389–405. https://doi.org/10.1177/0146621615569504

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*, 778–793. https://doi.org/10.1177/0013164408324460

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 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*, 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20. https://doi.org/10.3102/1076998607302626

van der Linden, W. J. (2009a). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25–41. https://doi.org/10.1177/0146621607314042

van der Linden, W. J. (2009b). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

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

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.

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*, 195–210. https://doi.org/10.1177/01466219922031329

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*, 418–438. https://doi.org/10.3102/1076998612466143

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*, 1017–1035. https://doi.org/10.1177/0013164413498256

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*, 95–109. https://doi.org/10.3758/s13428-011-0143-3

Wang, W.-C., & Qiu, X.-L. (2018). Multilevel modeling of cognitive diagnostic assessment: The multilevel DINA example. *Applied Psychological Measurement*, *43*, 34–50. https://doi.org/10.1177/0146621618765713

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 262–286. https://doi.org/10.1111/bmsp.12114

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*, 608–624. https://doi.org/10.1177/0146621616665196

Received 8 November 2016; revised version received 4 January 2019