JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Winter 2014, Vol. 51, No. 4, pp. 358–380

An Enhanced Approach to Combine Item Response Theory With Cognitive Diagnosis in Adaptive Testing

Chun Wang
University of Minnesota
Chanjin Zheng and Hua-Hua Chang
University of Illinois at Urbana-Champaign

Computerized adaptive testing offers the possibility of gaining information on both the overall ability and cognitive profile in a single assessment administration. Some algorithms aiming for these dual purposes have been proposed, including the shadow test approach, the dual information method (DIM), and the constraint weighted method. The current study proposed two new methods, aggregate ranked information index (ARI) and aggregate standardized information index (ASI), which appropriately addressed the noncompatibility issue inherent in the original DIM method. More flexible weighting schemes that put different emphasis on information about general ability (i.e., θ in item response theory) and information about cognitive profile (i.e., $\boldsymbol{\alpha}$ in cognitive diagnostic modeling) were also explored. Two simulation studies were carried out to investigate the effectiveness of the new methods and weighting schemes. Results showed that the new methods with the flexible weighting schemes could produce more accurate estimation of both overall ability and cognitive profile than the original DIM. Among them, the ASI with both empirical and theoretical weights is recommended, and attribute-level weighting scheme is preferred if some attributes are considered more important from a substantive perspective.

The new federal grant program entitled "Race to the Top" (RTTT) leads us into a new era of K-12 assessments in which both accountability and instructional improvement are emphasized (Chang, 2012). It reflects the growing importance of student assessment by putting an emphasis on developing statewide longitudinal data warehouses for monitoring student growth and learning, so that teachers can provide highly targeted and effective instruction in order to prepare the next generation of students for success in college and the workforce (U.S. Department of Education, 2009). Thus, in addition to providing a summary score for accountability purpose, providing diagnostic information to promote instructional improvement becomes an important goal of the next-generation assessment. This new mission has been reflected in the design of the Assessment of Readiness for College and Careers as well as the Smarter Balanced Consortium. In these two assessment systems, both a summative assessment component and a formative assessment component, are included (Partnerships for Assessment of Readiness for College and Careers, 2013; Smarter Balanced Assessment Consortium, 2013).

In light of this, we consider the problem of obtaining the estimation of a general ability (denoted by θ) as well as the diagnostic information on more specific skills (denoted by $\boldsymbol{\alpha}$) in a single administration of the computerized adaptive testing (CAT).

Designing a test that targets at both general ability estimation and specific cognitive feedback is not entirely new (Gibbons & Hedeker, 1992). K. K. Tatsuoka (1991) first proposed a rule-space methodology that provides a sound framework to show how to extract useful attribute mastery profile (i.e., cognitive information) from a test that is originally designed to produce a total score. Different from the original proposal by K. K. Tatsuoka (1991), we propose to use a model-based approach that replaces rule-space framework with more structured diagnostic classification models. In addition, we plan to deliver the test via adaptive testing mode such that the latent traits will be more effectively estimated. Presently, most research concerning item selection rules in CAT are based upon either item response theory (IRT) or cognitive diagnostic models (CDM) separately. We proposed to build a CAT in which the test is tailored interactively not only to each examinee's overall ability level, but also to each examinee's attribute mastery level, thereby information carried by both θ and $\boldsymbol{\alpha}$ will be maximized.

Three studies that addressed the dual-purpose item selection in adaptive testing are McGlohen and Chang (2008), Cheng and Chang (2007), and Wang, Chang, and Douglas (2012). They solved the dual-objective optimization problem through three different strategies. McGlohen and Chang (2008) proposed a two-stage method, in which the "shadow" test functions as a bridge to connect information gathered at θ for IRT and information accumulated at $\boldsymbol{\alpha}$ for CDM. Wang et al. (2012) proposed a constraint-weighted item selection algorithm that treats information at θ as the objective function and information at $\boldsymbol{\alpha}$ as the statistical constraints. Cheng and Chang (2007) proposed a dual information index which is a weighted sum of information gathered at θ and $\boldsymbol{\alpha}$, respectively. In this paper, we extend the dual information index in the following ways: first, the two information pieces are rescaled to make them comparable; second, the weights are selected either based on theoretical findings or empirical needs; finally, the solution for dealing with the case in which distinctive attributes are treated differently is proposed.

The remainder of the paper is organized as follows. First we introduce the CDMs and discuss the situations under which both the diagnostic models and unidimensional IRT model can fit simultaneously to the same data. We then introduce the original dual information method (DIM) developed by Cheng and Chang (2007), followed by the two new item selection algorithms—aggregate ranked information (ARI) index and aggregate standardized information (ASI) index. Then three different schemes of selecting weights are proposed. Two simulation studies are conducted to examine the performance of the new methods. Discussions and practical guidelines are given in the end.

Psychometric Models

In the current research, the DINA model (deterministic inputs, noisy "and" gate; Junker & Sijtsma, 2001) is used for cognitive diagnostics. Let Y_{ij} be the response of examinee i to item j, $i = 1, \ldots, I$, $j = 1, \ldots, J$, and let $\boldsymbol{\alpha}_i = \{\boldsymbol{\alpha}_{ik}\}$ be the examinee's skill vector, $k = 1, \ldots, K$, where $\boldsymbol{\alpha}_{ik} = 1$ denotes mastery of skill k and 0 otherwise. A Q-matrix (Embretson, 1984; K. K. Tatsuoka, 1985), which is a $J \times K$ matrix with 0s and 1s as its entries, explicitly identifies the cognitive specification for each item. The element in the jth row and kth column of the matrix, q_{jk}, indicates whether skill

k is required to correctly answer item j. An examinee's skill vector and the Q-matrix produce, in a conjunctive manner, an "ideal" response vector $\eta_i = \{\eta_{ij}\}$where

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}} = \begin{cases} 1 & \text{if examinee } i \text{ possesses all the required skills for item } j \\ 0 & \text{if examinee } i \text{ lack of at least one of the required skills for item } j \end{cases}.$$

The slipping and guessing parameters of item j are defined as $s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$, respectively. Therefore, the probability of examinee i with the skill vector $\boldsymbol{\alpha}_i$ answering item j correctly can be defined as

$$P_j(\boldsymbol{\alpha}_i) = P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}.$$

It is worthwhile to point out that the choice of the DINA model in the current study is for demonstration purpose and the proposed methods can be generalized to other diagnostic classification models such as the NIDA (noisy inputs, deterministic "and" gate; Maris, 1999), and the reparameterized unified models (de la Torre & Douglas, 2004; Hartz, 2002) , among others.

Aside from the DINA model for $\boldsymbol{\alpha}$, we assume the overall ability θ arises from a separate item response model (i.e., 2PL/3PL model). De la Torre and Douglas (2004) proposed a higher-order DINA (HO-DINA) model that combines the IRT model and diagnostic model by assuming conditional independence of response Y given $\boldsymbol{\alpha}$, and also assuming that the components of $\boldsymbol{\alpha}$ are independent conditional on θ. Their approach can parsimoniously model the joint distribution of the attributes and meanwhile estimate both θ and $\boldsymbol{\alpha}$ simultaneously through one calibration. The particular relationship between θ and $\boldsymbol{\alpha}$ in their paper is logistic regression given as

$$P(\boldsymbol{\alpha}_k = 1\,|\theta) = \frac{\exp(\lambda_{0k} + \lambda_k\theta)}{1 + \exp(\lambda_{0k} + \lambda_k\theta)}. \tag{1}$$

The correlation between attributes is determined by the size of slope λ_k and intercept λ_{0k}. Even though the HO-DINA model provides a new avenue for dual-purpose tests, we have to emphasize that the current paper is built upon the assumption that we can fit the same data set with both the unidimensional IRT model and the CDM separately, so as to gain two sets of item/person parameters for the dual purposes. The justification will be expounded upon in the follow-up sections.

Fitting Separate Models Is Practically Preferable in Some Cases

Although de la Torre and Douglas's (2004) integrated HO-DINA model has received much attention recently (it has been cited over 200 times already based on Google Scholar citations), we believe one profound advantage of fitting separate models is that estimating θ from well-calibrated item parameters allows linking responses from different test forms, or multiple groups, or multiple time points, on the same scale using well-established linking/equating methods. For instance, if one needs to track students' growth on θ over years, only when θ estimates from different time points are put on the same scale that can one gauge the growth properly. Using the HO-DINA model, if different sets of attributes are measured in different years, or if the loading structure (i.e., λ_k and λ_{0k}) changes across time points, the measurement invariance of "overall ability" will be hard to justify, and consequently

it will be difficult to evaluate students' growth on the general ability. Using separate modeling approach, on the other hand, if the anchor item parameters stay the same over time, the measurement invariance of θ is automatically satisfied (Grimm, Kuhl, & Zhang, 2013).

According to McGlohen and Chang (2008) and H. Liu, You, Wang, Ding, and Chang (2012), two separate models can work effectively on certain types of large-scale assessment. This assumption is further defended via a small-scale simulation study presented in Appendix A. From psychometric point of view, the conditions where both the unidimensional IRT model and the DINA model can be fitted properly with the same data set are (1) the attributes measured by the test are highly correlated (see Appendix A for further evidence); and (2) the attributes display a linear hierarchical relationship (Leighton & Gierl, 2007; von Davier, 2013; von Davier & Haberman, 2014), meaning that mastering higher-level skills requires mastery of all lower-level skills. For instance, if $K = 3$, then the number of all possible skill patterns is 4 (i.e., $\boldsymbol{\alpha} = [0,0,0], [1,0,0], [1,1,0], [1,1,1]$) rather than 8. Von Davier and Haberman (2014) named such a structure a *deterministic unidimensional order* because they claimed that the linear hierarchy in latent profiles form a perfect Guttman pattern (Guttman, 1950).

Practical Guidelines

In real data applications, if practitioners decide to take the separate model-fitting approach suggested in this paper, there are several specific steps to take. First, they need to have pre-knowledge and substantive understanding of the content of the tests such as whether the attributes being measured are highly correlated, or exhibit a linear structure. Second, they can fit both the diagnostic classification models and the unidimensional IRT models to the response data and check both the model and item level fit. Interested readers are referred to Chen, de la Torre, and Zhang (2013) for available model checking methods for DCMs, and Stone and Zhang (2003) for IRT model fit evaluation. This model fit evaluation guideline is executed in the simulation study (see Appendix B for details) showing that the two separate models provide adequate fit to a data set generated from the HO-DINA model with high correlation among the attributes.

Item Selection Methods

Once both psychometric models are assumed to fit the data set appropriately, the next challenge is to define an item selection method. Denote θ as the unidimensional or multidimensional continuous latent ability the test intends to measure. The Kullback-Leibler information of θ, first introduced by Chang and Ying (1996) in adaptive item selection, is defined as

$$K(\hat{\theta} \,\|\theta) = \sum_{y=0}^{1} \log \left(\frac{P(Y_j = y|\hat{\theta})}{P(Y_j = y|\theta)} \right) P(Y_j = y|\hat{\theta}), \tag{2}$$

where $\hat{\theta}$ is the intermediate ability estimate and Y_j denotes the response of a given examinee on item j. $P(Y_j = y\,|\theta)$ is the item response function given an IRT model.

They further propose an item selection criterion, the KL information index (KI), expressed as

$$KL(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} K(\hat{\theta} \,\|\theta)d\theta, \tag{3}$$

where δ is usually chosen to be $\frac{3}{\sqrt{n}}$ with n being the number of items administered so far.

Further assume the test measures K specific attributes denoted by a latent vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$. For a math ability test, the specific attributes could be subtraction, fraction, multiplication, etc. The appropriate KL information index would thus calculate the divergence between the conditional distribution of Y_j given the provisional estimated state, $f(Y_j|\hat{\boldsymbol{\alpha}})$, and the conditional distribution of Y_j given the true latent state, $f(Y_j|\boldsymbol{\alpha})$, which is computed as

$$KL_j(\hat{\boldsymbol{\alpha}} \,\|\boldsymbol{\alpha}) = \sum_{y=0}^{1} \log\left(\frac{P(Y_j = y\,|\hat{\boldsymbol{\alpha}})}{P(Y_j = y\,|\boldsymbol{\alpha})}\right) P(Y_j = y\,|\hat{\boldsymbol{\alpha}}). \tag{4}$$

Because the true latent profile, $\boldsymbol{\alpha}$, is generally unknown, one cannot directly calculate (4). Xu, Chang, and Douglas (2003) propose to sum up the KL divergences between $f(Y_j|\hat{\boldsymbol{\alpha}})$ and the conditional distribution of Y_j given each possible latent profile. Their KL index can be written as

$$KL_j(\hat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} \left[\sum_{y=0}^{1} \log\left[\frac{P(Y_j = y\,|\hat{\boldsymbol{\alpha}})}{P(Y_j = y\,|\boldsymbol{\alpha}_c)}\right] P(Y_j = y\,|\hat{\boldsymbol{\alpha}})\right], \tag{5}$$

where c indexes attribute profile and runs from 1 to 2^K. Recently, Cheng (2009) propose a posterior weighted KL information index (PWKL). The idea is to multiply each addend, $KL_j(\hat{\boldsymbol{\alpha}} \,\|\boldsymbol{\alpha}_c)$, by its posterior density given response pattern on the previously selected items, as follows:

$$PWKL_j(\hat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} P(\boldsymbol{\alpha}_c|y) \left[\sum_{y=0}^{1} \log\left[\frac{P(Y_j = y\,|\hat{\boldsymbol{\alpha}})}{P(Y_j = y\,|\boldsymbol{\alpha}_c)}\right] P(Y_j = y\,|\hat{\boldsymbol{\alpha}})\right]. \tag{6}$$

In (6), $P(\boldsymbol{\alpha}_c|\mathbf{y})$ is the posterior density of $\boldsymbol{\alpha}_c$ given an examinee's response vector, $\mathbf{y}$, on the previously administered items. Cheng (2009) has shown that selecting items by maximizing PWKL information yielded higher measurement precision as compared to alternative methods. Wang (2013) later proposed a mutual information method, which is shown to be very efficient when test length is short.

The problem of building a CAT to accurately estimate both θ and $\boldsymbol{\alpha}$ can be treated as a dual-objective optimization problem; the two objective functions are information gathered at θ and information accumulated at $\boldsymbol{\alpha}$, respectively. A general way of dealing with dual-objective optimization problem is to reduce it to a single-objective problem through some aggregation techniques, and we will focus on this approach in this study. Because the information pieces for $\boldsymbol{\alpha}$ and θ will be put together in an aggregate index, it is more intuitive to use the same type of information criterion, in this case, the KL information, for both $\boldsymbol{\alpha}$ and θ.

Constructing a Single Aggregate Objective Function

The basic idea is to combine both objective functions into a single functional form. A straightforward combination is a weighted linear sum of the objectives, defined as follows:

$$Objective = wPWKL(\hat{\boldsymbol{\alpha}}) + (1 - w)KL(\hat{\theta}), \tag{7}$$

where w is the weight, and $\hat{\boldsymbol{\alpha}}$ and $\hat{\theta}$ are the intermediate estimates after each item is administered. This DIM is first proposed by Cheng and Chang (2007). We address two issues innate in their original method: noncomparability of the two information addends and the arbitrary selection of weight. Notice that even though the PWKL index (Chang & Ying, 1996) is considered in (7) and in all arguments hereafter, the two new methods that will be introduced below allow any form of information criterion, including the mutual information index (Wang, 2013), or likelihood weighted KL index (Barrada, Olea, Ponsoda, & Abad, 2008), or Shannon entropy index (Xu et al., 2003), in forming the aggregate item selection index.

The noncomparability appears as a result of integration. Specifically, the integral in (3) is transformed to the summation as follows:

$$\begin{aligned} KL(\hat{\theta}) &= \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \left[p(\hat{\theta}) \log\left(\frac{p(\hat{\theta})}{p(\theta)}\right) + (1 - p(\hat{\theta})) \log\left(\frac{1 - p(\hat{\theta})}{1 - p(\theta)}\right)\right] d\theta \\ &\approx \sum_{i=-t}^{t} \left[p(\hat{\theta}) \log\left(\frac{p(\hat{\theta})}{p(\hat{\theta} + i\,\Delta\theta)}\right) + (1 - p(\hat{\theta})) \log\left(\frac{1 - p(\hat{\theta})}{1 - p(\hat{\theta} + i\,\Delta\theta)}\right)\right] \Delta\theta. \end{aligned} \tag{8}$$

From (3) and (5), it is clear that $KL(\hat{\boldsymbol{\alpha}})$is comprised of 2^K addends, whereas the number of addends in $KL(\hat{\theta})$ depends on how you slice the integration domain—"t." In addition, the size of the terms in (7) differs greatly because every addend in (8) has $\Delta\theta$ as a multiplier, which means $KL(\hat{\theta})$ will always be smaller than $KL(\hat{\boldsymbol{\alpha}})$. Therefore $KL(\hat{\boldsymbol{\alpha}})$will play a dominant role in item selection if (7) is considered.

Figure 1 lends further assertions to the incompatibility of two information pieces in (7). It shows the scatter plot of these two information pieces for each item in the item bank, given intermediate $\hat{\theta}$ and $\hat{\boldsymbol{\alpha}}$ estimations. Each star in the scatter represents an item, KL information on $\hat{\boldsymbol{\alpha}}$ is much larger than that on $\hat{\theta}$. Items fall in the ellipse are more preferable because they provide higher information on both $\hat{\theta}$ and $\hat{\boldsymbol{\alpha}}$. However, due to the fact that $KL(\hat{\boldsymbol{\alpha}})$ is large (please see the scale of x-axis as opposed to the scale of the y-axis) that items within the rectangular will also be preferred although some of them really provide little information on $\hat{\theta}$, whereas the items within purple circle provides more balanced information on both $\hat{\theta}$ and $\hat{\boldsymbol{\alpha}}$. To solve this non-comparability issue, we propose the following two modifications.

Aggregate ranked information method (ARI). The idea is simply to transform both pieces of information to an ordinal scale in such a way that each item will have two ranks for $KL(\hat{\theta})$ and $PWKL(\hat{\boldsymbol{\alpha}})$ separately. ARI is therefore computed as

$$ARI = wpe(PWKL(\hat{\boldsymbol{\alpha}})) + (1 - w)pe(KL(\hat{\theta})), \tag{9}$$

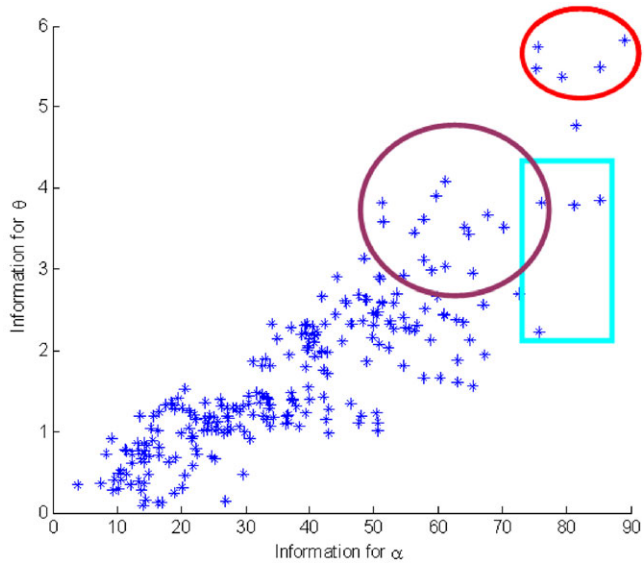

Figure 1. Illustration of KL information for $\hat{\theta}$ and $\hat{\boldsymbol{\alpha}}$.

where $pe(\bullet)$ represents "rank." The rationale behind this method is that by using an ordinal scale, the information captured by θ and $\boldsymbol{\alpha}$ can be aligned along the same scale and the weight w $(0 \leq w \leq 1)$ will reflect the true relative importance of the two pieces.

Aggregate standardized information method (ASI). When transforming from the original ratio scale to an ordinal scale in ARI, some information will be lost. It would be better to keep the ratio scale and resort to scale transformation. In our case, notice that both $KL(\hat{\theta})$ and $PWKL(\hat{\boldsymbol{\alpha}})$are summations of KL information of two probabilities, the key differences being the number of addends and the size of each addend. Therefore, we can standardize both of them to remove the scale difference

$$KL^*(\hat{\theta}) = \frac{(KL(\hat{\theta}) - mean(KL(\hat{\theta})))}{SD(KL(\hat{\theta}))},$$
$$PWKL^*(\hat{\boldsymbol{\alpha}}) = \frac{PWKL(\hat{\boldsymbol{\alpha}}) - mean(PWKL(\hat{\boldsymbol{\alpha}}))}{SD(PWKL(\hat{\boldsymbol{\alpha}}))}, \tag{10}$$

$$ASI = wPWKL^*(\hat{\boldsymbol{\alpha}}) + (1 - w)KL^*(\hat{\theta}). \tag{11}$$

In real applications when exposure control and content balancing are of concern, off-the-shelf methods, such as the Sympson-Hetter (Sympson & Hetter, 1985) method, or restrictive stochastic methods (Wang, Chang, & Huebner, 2011) for exposure control, as well as the maximum priority index (Cheng & Chang, 2009) for content balancing, can be applied. One simply needs to replace the Fisher

information index, or other information-based criterion in the above-mentioned methods by either ASI or ARI.

Selection of Weights

Clearly, in the aggregate approach, the solution obtained will depend on the values of the specified weight w. In Cheng and Chang (2007)'s method, they chose 11 different arbitrary values of w, from 0 to 1 with .1 interval; however, in this study, we propose three additional weighting schemes.

Theory-based weights. The binary ability vector $\boldsymbol{\alpha}$ can be regarded as finite partially ordered set (C. Tatsuoka, 2002), and various CDMs belong to discrete poset classification models. Item selection rules for either purely IRT-based CAT or diagnostic classification model-based CAT is to find a set of items so that $\hat{\theta}$ converges to θ ($\hat{\boldsymbol{\alpha}}$ converges to $\boldsymbol{\alpha}$) as fast as possible. Ideally, the optimal rates of convergence for discrete poset classification models are exponential (C. Tatsuoka, 2002), whereas the optimal convergence rate for $\hat{\theta}$ is $\sqrt{n}$ (n is the number of items been administered; Chang & Stout, 1993). Therefore, the $\hat{\boldsymbol{\alpha}}$ converges to $\boldsymbol{\alpha}$ faster than $\hat{\theta}$ converges to θ, and it is reasonable to give more weight to $KL(\hat{\boldsymbol{\alpha}})$at the beginning of the test to accelerate its convergence. At the later stage of the test when $\hat{\boldsymbol{\alpha}}$ is estimated accurately, more weight can be put on $KL(\hat{\theta})$. A simple way to define weight to reflect this transition is $w = 1 - n/L$ where n is the number of items that have been chosen so far, and L is the test length. This approach assumes that the test length L is determined in advance, which is the case with a fixed-length CAT. When a variable-length CAT is considered, this weighting scheme cannot be applied directly.

Empirical weights. The weights can also be chosen empirically in that we try to balance the contribution of both information pieces—whenever one information piece lags behind, we assign more weight to it. Such an idea is reflected by the following definition of weight w:

$$\left.\begin{aligned} w_1 &= (u_\theta - x_\theta^{(k)})/u_\theta \\ w_2 &= (u_{\boldsymbol{\alpha}} - x_{\boldsymbol{\alpha}}^{(k)})/u_{\boldsymbol{\alpha}} \end{aligned}\right\} w = \frac{w_2}{w_1 + w_2}, \tag{12}$$

where u_θ and $u_{\boldsymbol{\alpha}}$ are the pre-chosen upper bounds of the total information at θ and $\boldsymbol{\alpha}$, respectively, and $x_\theta^{(k)}$ and $x_{\boldsymbol{\alpha}}^{(k)}$ are the accumulated information at θ and $\boldsymbol{\alpha}$ after k items have been administered (i.e., $x_\theta^{(k)} = \sum_{j=1}^{k} PWKL_j(\hat{\theta})$, and $x_{\boldsymbol{\alpha}}^{(k)} = \sum_{j=1}^{k} KL_j(\hat{\boldsymbol{\alpha}})$).

The upper bound is often determined in the following way. For instance, to determine u_θ, researchers can plot the distribution of $\sum_{j=1}^{J} KL_j(\hat{\theta})$ (i.e., sum of the KL information over J (test length) most informative items) at a range of different $\hat{\theta}$ values, and pick the highest value as u_θ . Same procedures can be applied to find $u_{\boldsymbol{\alpha}}$. For instance, Figure 2 shows the test KL information for a 20-item test from a 500-item pool (from our simulation study 1 below). Figures 2(a) and 2(b) show the KL information for θ and $\boldsymbol{\alpha}$, respectively, whereas Figures 2(c) and 2(d) present the PWKL information for $\boldsymbol{\alpha}$ (i.e., Equation 6) with either a hypothetically uniform posterior or a nonuniform posterior density of $\boldsymbol{\alpha}$. In this case, if $KL_j(\hat{\boldsymbol{\alpha}})$ and $KL_j(\theta)$ are considered, then the upper bound u_θ can be set at 3, and the upper bound $u_{\boldsymbol{\alpha}}$ can be fixed at

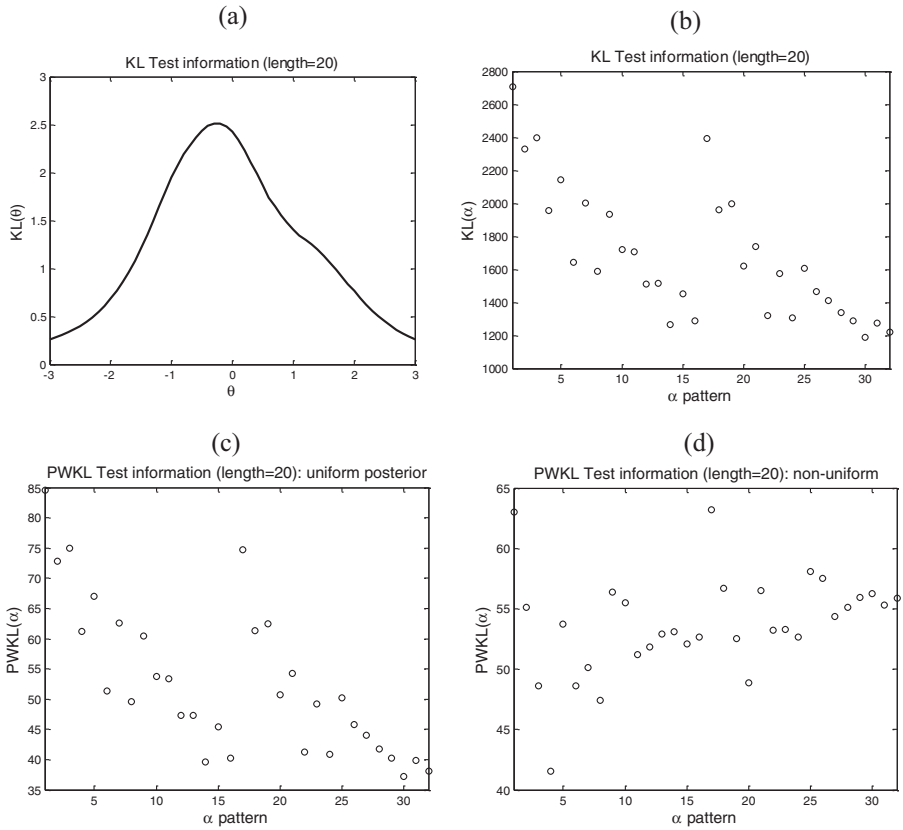

Figure 2. Test KL information for a 20-item test.

3,000. Otherwise, if $PWKL_j(\hat{\boldsymbol{\alpha}})$ is considered, then the upper bound u_α can be set 90. As shown in this illustration, the selection of upper bounds sometimes needs to be tweaked on a case-by-case basis; it is largely affected by the quality of the items in the item bank. Moreover, adding the posterior weight in $PWKL_j(\hat{\boldsymbol{\alpha}})$, to some extent, mitigates the incompatibility issues inherent in the original DIM because the scale difference between $PWKL_j(\hat{\boldsymbol{\alpha}})$ and $KL_j(\theta)$ is reduced.

The weight defined here has a built-in "minimax mechanism"—it tends to pick the items that maximize the information of the estimator (either $\hat{\theta}$ or $\hat{\boldsymbol{\alpha}}$) lagging behind. In other words, using the weight defined in (9) in either ARI or ASI will tend to select items with large information for the estimator with a larger gap between the information gathered and the upper bound. Consequently, information accumulated at both θ and $\boldsymbol{\alpha}$ can approximate their upper bounds as closely as possible.

Attribute-level weights. When a test measures multiple traits (or attributes), it is often the case that some traits are more important than the others, such as the distinction between intentional abilities and nuisance abilities (van der Linden, 1996;

Veldkamp & van der Linden, 2002). This relative importance of different attributes can be reflected in the construction of $KL_j(\hat{\boldsymbol{\alpha}})$ or $PWKL_j(\hat{\boldsymbol{\alpha}})$.

If we denote $D_{juv} = \sum_{y=0}^{1} \log[\frac{P(Y_{ij}=y|\boldsymbol{\alpha}_u)}{P(Y_{ij}=y|\boldsymbol{\alpha}_v)}] P(Y_{ij} = y\,|\boldsymbol{\alpha}_u)$ as the discrimination power of item j in terms of differentiating item response function generated by $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$, assuming the current intermediate estimate is $\boldsymbol{\alpha}_u$, then the KL information index can be computed as

$$KL_j(\boldsymbol{\alpha}_u) = [w_1, w_2, ...w_K] \begin{bmatrix} \sum_{\substack{v=1 \\ \boldsymbol{\alpha}_v \neq \boldsymbol{\alpha}_u}}^{2^K} d_1(u,v) D_{juv} \\ \vdots \\ \sum_{\substack{v=1 \\ \boldsymbol{\alpha}_v \neq \boldsymbol{\alpha}_u}}^{2^K} d_M(u,v) D_{juv} \end{bmatrix}, \tag{13}$$

where $w_1, w_2, \ldots w_K$ are the user-defined weights that reflect the relative importance of each attribute and $\sum_{i=1}^{K} w_i = 1$; $d_i(u,v), i = 1, 2, \ldots K$ is defined as the number of different attributes between $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$:

$$d_i(u,v) = \begin{cases} \frac{1}{d} & \text{if } \boldsymbol{\alpha}_{ui} \neq \boldsymbol{\alpha}_{vi} \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

where d is the total number of different attributes between $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$. For example, if $\boldsymbol{\alpha}_u = [1, 0, 0, 0, 1]$ and $\boldsymbol{\alpha}_v = [0, 1, 1, 0, 1]$, $d_i(u,v) = \frac{1}{3}, i = 1, 2, 3$, and $d_i(u,v) = 0, i = 4, 5$; therefore D_{juv} will not enter into the summation when calculating the information for the fourth and fifth attributes. The rationale behind this treatment is that the component D_{juv} contributes equally to the sets of attributes that differ by $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$; by doing so, we are able to differentiate the "information" contribution of item to each and every attribute. The weighting scheme introduced in this subsection is different from the previous section in that the weight is assigned to each attribute when constructing the KL information piece on $\hat{\boldsymbol{\alpha}}$. Note that the posterior density of each possible pattern, $P(\boldsymbol{\alpha}_v|y)$, which is included in the PWKL index, can also be considered in this weighting scheme such that (13) is updated as follows:

$$KL_j(\boldsymbol{\alpha}_u) = [w_1, w_2, ...w_M] \begin{bmatrix} \sum_{\substack{v=1 \\ \boldsymbol{\alpha}_v \neq \boldsymbol{\alpha}_u}}^{2^M} d_1(u,v) D_{juv} P(\boldsymbol{\alpha}_v|\mathbf{y}) \\ \vdots \\ \sum_{\substack{v=1 \\ \boldsymbol{\alpha}_v \neq \boldsymbol{\alpha}_u}}^{2^M} d_M(u,v) D_{juv} P(\boldsymbol{\alpha}_v|\mathbf{y}) \end{bmatrix}. \tag{15}$$

Simulation Studies

Two simulation studies were conducted to check the proposed two aggregate indices and three different weighting schemes. In both studies, θ and $\boldsymbol{\alpha}$ were estimated

intermediately during the entire course of the adaptive test. $\boldsymbol{\alpha}$ was estimated by maximum likelihood estimation, i.e., the pattern that generates the highest likelihood is $\hat{\boldsymbol{\alpha}}$ (Huebner & Wang, 2011). θ was estimated by both MLE and expected a posterior (EAP). When an examinee answered all current items correctly or incorrectly, EAP was used; otherwise, MLE was used (Chang & Ying, 1999). For each simulation study, we had a complete crossed design resulting in: three-item selection methods (original DIM, ARI, ASI) $\times$ four weighting schemes (arbitrary grid in which the weight is selected to be 0, .5, and 1; theoretical; empirical; attribute-level weight) $\times$ two test lengths (short versus long) $=$24 conditions.

Study 1

This study was conducted assuming the number of attributes is 5, a medium number that is often considered in the literature (Wang, 2013). A 500-item bank was generated with a 500-by-5 Q-matrix. The number of items measuring each attribute was balanced, ranging from 234 to 257. The slipping and guessing parameters were then simulated from a 4-Beta distribution following de la Torre and Douglas (2004). A total of 3,000 examinees were simulated from the higher-order DINA model. Specifically, 3,000 higher-order θ_0's were first drawn from a standard normal N (0, 1) distribution. The slope and intercept parameters in the HO-DINA model were then chosen such that the resulting correlations among the attributes were between .45 and .65, similar to the correlations obtained from the operational CAT-ASVAB (Armed Services Vocational Aptitude Battery) test (Segall, 1996). Examinees' mastery profiles were generated by comparing $P(\boldsymbol{\alpha}_k = 1 \,|\theta)$ with a random value, u, obtained from a uniform (0, 1) distribution, and $\boldsymbol{\alpha}_k = 1$ if $P(\boldsymbol{\alpha}_k = 1 \,|\theta) \geq u$. A 3000-by-500 complete response matrix was generated based on the HO-DINA model, and it was retrofitted with the 2PL model using BILOG and with the DINA model using the EM algorithm (de la Torre, 2009). As a result, we obtained a-(discrimination) and b-(difficulty), slipping, and guessing parameters for the same 500-item pool through retrofitting, and the corresponding θs and $\boldsymbol{\alpha}$s were treated as examinees' "true" abilities and "true" cognitive profiles. The descriptive statistics of the item pool and examinee sample are presented in Tables C1 and C2 in Appendix C. Examinee responses to each item in CAT were generated from the DINA model based on their interim cognitive pattern estimates. Two criteria are presented to evaluate the performance of the three algorithms: the attribute recovery rate and the pattern recovery rate (Wang et al., 2011) for the cognitive profile $\boldsymbol{\alpha}$ and root mean squared error (RMSE) for the IRT ability parameter.

Results. Table 1 presents the results for the combinations of three different item selection algorithms and three weighting schemes (including grid weights, theoretical weights, and empirical weights) under two different test length conditions. Table 2 presents the results for attribute-level weighting method (against balanced weighting method). The weight vector, w, was chosen to be (.05, .05, .05, .8, .05), to intentionally up-weight the fourth attribute.

Several conclusions can be drawn from the results. First, increasing test length increased the recovery rates of $\boldsymbol{\alpha}$ yet not necessarily decreased MSE of θ estimation. This is because in our simulation, item responses were generated based on $\boldsymbol{\alpha}$, such

Table 1
Results of ARI, ASI, and Original DIM for 5 Attributes

		Pattern Recovery			RMSE		
Test Length	Weights	DIM	ARI	ASI	DIM	ARI	ASI
20	0[a]	.755	.757	.757	.354	.353	.354
	.5	.931	.965	.973	.258	.305	.27
	1	.978	.978	.979	.564	.563	.561
	Theoretical	.938	.98	.982	.248	.364	.236
	Empirical	.996	.98	.984	.322	.318	.263
30	0	.897	.899	.898	.431	.433	.433
	.5	.984	.983	.991	.281	.343	.263
	1	.995	.995	1	.356	.356	.356
	Theoretical	.959	.98	1	.248	.302	.236
	Empirical	.994	1	1	.311	.267	.263

[a]When the weights are 0 or 1, ARI, ASI, and DIM generated essentially the same results. This is because when only information on θ (or $\boldsymbol{\alpha}$) is used for item selection, monotonic transformation in ARI and ASI does not change the item order. So the three methods basically selected the same set of items with the largest information. The observed tiny difference in Table 1, even though negligible, is due to the randomness in generating item responses in our simulation study.

Table 2
Comparisons Between Attribute-Level Weight and Balanced Weight

Test Length	Method	Weights	At1	At2	At3	At4	At5	Pattern
20	ARI	.5	1.000	.998	.989	.964	.966	.920
		Attribute	1.000	.991	.989	.988	.981	.957
	ASI	.5	1.000	1.000	.992	.917	.963	.877
		Attribute	.997	1.000	.995	1.000	.909	.904
	DIM	.5	.999	.999	.999	.984	.997	.978
		Attribute	.990	.992	.998	1.000	.952	.934
30	ARI	.5	1.000	.999	.997	.967	.962	.928
		Attribute	.999	.995	.989	.991	.985	.958
	ASI	.5	1.000	1.000	.999	.966	.994	.959
		Attribute	.999	1.000	.998	1.000	.951	.947
	DIM	.5	1.000	1.000	1.000	.989	.997	.986
		Attribute	.998	.996	.999	1.000	.985	.978

Note. The baseline results where the weight = .5 were obtained when KL information index was used rather than the PWKL index. The attribute level weighting was based on (13).

that θ is not directly related to the actual responses. This also explains why setting the weight = 0 does not necessarily improve θ estimation because accurately estimating θ relies on correctly recovering $\boldsymbol{\alpha}$. Second, both ARI and ASI generated more balanced results than the original DIM, with higher pattern recovery rates of $\boldsymbol{\alpha}$ but slighted increased MSE of θ. Third, both the empirical and theoretic weighting methods generated more accurate estimates of θ and $\boldsymbol{\alpha}$ (with ARI and ASI) as opposed

Table 3
Results of ARI, ASI, and Original DIM for 8 Attributes

		Pattern Recovery			RMSE		
Test Length	Weights	DIM	ARI	ASI	DIM	ARI	ASI
36	0	.321	.322	.324	.592	.595	.584
	.5	.834	.664	.843	.408	.366	.406
	1	.874	.882	.883	.469	.450	.477
	Theoretical	.711	.555	.726	.394	.379	.425
	Empirical	.868	.763	.872	.401	.384	.418
60	0	.421	.431	.421	.420	.411	.408
	.5	.936	.825	.956	.285	.261	.315
	1	.963	.970	.971	.363	.383	.378
	Theoretical	.891	.692	.923	.309	.292	.313
	Empirical	.964	.906	.970	.303	.281	.332

Table 4
Comparisons Between Attribute-Level Weight and Balanced Weight

Test Length	Method	Weights	At1	At2	At3	At4	At5	AT6	At7	At8	Pattern
36	ARI	.5	.991	.949	.987	.967	.921	.874	.960	.955	.664
		Attribute	.978	.935	.989	.965	.930	.906	.958	.945	.665
	ASI	.5	.995	.988	.990	.984	.949	.981	.962	.974	.843
		Attribute	1	.975	.990	.985	.980	.980	.970	.975	.875
	DIM	.5	.998	.991	.993	.986	.949	.971	.967	.961	.834
		Attribute	.994	.970	.993	.982	.967	.980	.972	.973	.848
60	ARI	.5	.999	.982	.999	.991	.958	.917	.991	.974	.825
		Attribute	.999	.977	.997	.987	.948	.933	.991	.976	.823
	ASI	.5	.999	1	.999	.997	.992	.993	.992	.986	.956
		Attribute	1	1	.998	.998	.994	.988	.996	.992	.968
	DIM	.5	.999	.994	.999	.994	.980	.992	.993	.986	.936
		Attribute	.999	.990	.999	.997	.989	.992	.994	.992	.952

to the balanced weights. Lastly, because it was observed that the fourth attribute had lowest recovery rate among the five attributes, the attribute-level weighting method assigned more weight to Attribute 4 and, as a result, produced a better attribute recovery rate for Attribute 4 in Table 2. However, increasing attribute recovery rate sometimes sacrificed the pattern recovery rate (e.g., see the DIM with test length = 20 or 30) because some attributes are inevitably down-weighted, yielding lower recovery rates for those items, which in turn affects the pattern recovery. Note that the baseline method in Table 2 was based on the KL index in (5) rather than the PWKL index because using PWKL, all attribute-level recovery rates are almost 1 when test length is 30 and varying attribute-level weights does not produce any visible difference.

Study 2

The second study differs from the first one in that both the 3PL and the DINA item parameters were obtained from a large-scale English language proficiency test, and the number of attributes equals to 8, a larger number that makes the estimation more difficult. Due to this large number of attributes, the test lengths considered in this study are 36 and 60.

Data description. Both the 3PL model and the DINA model (the Q-matrix was identified by content experts[1]) were fitted to a real response data (H. Liu et al., 2013), and the resulting item parameters were used in this study. Table C3 presents the summary statistics for the item parameters. The mean difficulty parameter was around zero and its standard deviation was large, indicating that the entire item bank had median difficulty level and the items had enough variability in terms of difficulty. According to Table C4, majority of the items measured one attribute and a few items measured two attributes. The number of items measuring each attribute was relatively even assuring the item bank had enough coverage for each attribute. The number of examinees mastering certain number of attributes indicated the sample had good coverage of examinees with different skill patterns and more than half of the examinees mastered four or more attributes.

Results. Table 3 presents the summary results for both the estimation of θ and $\boldsymbol{\alpha}$. Table 4 presents the results for attribute-level weighting method (against balanced weighting method). The weight vector, w, was chosen to be (.01, .02, .01, .01, .02, .90, .01, .01) for ARI, (.09, .09, .09, .11, .23, .11, .14, .14) for ASI, and (.09, .09, .09, .10, .23, .11, .16, .14) for DIM. The weights were selected based on our pilot study, and the primary rationale is that attributes with lower recovery rates were assigned higher weights.

The general pattern is the same as in Study 1. However, there are several interesting observations that merit comments. First, when the weight $= 0$, the recovery of both θ and α are the worst because, as we have emphasized, the recovery of θ heavily depends on the recovery of α. Second, it seems that ARI lost its advantage as compared to Study 1. This might be because the distribution of item information is rather skewed, whereas the transformed ranked information follows a uniform distribution ignoring the actual size of the information carried by each item, thus it is not as efficient as ASI. Third, theoretical weights no longer generated better results because the convergence of $\hat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}$ might be slow if $\boldsymbol{\alpha}$ is high-dimensional and when slipping and guessing parameters are relatively large. We also present the results for the attribute-level weighting for each method. Attributes with comparatively higher weights (e.g., attribute 5 in ASI method or attribute 6 in ARI method) were shown to have higher attribute-level recovery rate when using the attribute-level weighting method, which is consistent with our expectation. The entire pattern recovery rate was also improved as a result.

Summary

Based on the two simulation studies, we can conclude that both ASI and ARI improve upon the original DIM method by better balancing between information

of θ and information of $\boldsymbol{\alpha}$. In general, ASI should always be preferred over ARI because ARI loses efficiency when transforming a ratio scale information index to ordinal scale ranks. Moreover, both empirical and theoretical weights outperform the equal weight in most conditions, except when the slipping and guessing parameters are relatively large, in which case only empirical weight beats equal weight. As a result, the theoretical weight is recommended for high-quality item bank (due to its simplicity) whereas empirical weight is recommended for less informative item bank. In the end, the attribute-level weights can successfully up-weight the more important attributes, yet sometimes at the cost of worsening the recovery of other attributes.

Discussion

Most current CAT systems were originally developed for large-scale, high-stakes admission exams in which accurately estimating the total true score is the major concern. In K-12 assessment, on the other hand, teachers are also interested in getting instructional feedback. New item selection algorithms are needed to satisfy the dual purposes in educational settings. This study proposes practical methods that capitalize on the availability of cognitive diagnostic information from computerized adaptive tests and offer a possibility of transiting from the traditional CAT to a dual-purpose CD-CAT at a minimum cost.

Simulation results showed that by rescaling the two information pieces on θ and $\boldsymbol{\alpha}$, respectively, both ARI and ASI can solve the incomparability issue in DIM and provide better balance between the summative and the diagnostic information. The difference between ARI and ASI boils down to how the two information pieces are rescaled. ASI essentially transforms each information piece to a standardized Z-score while still maintaining the original ratio scale so that no information is lost for item selection purposes. On the other hand, ARI relies only on the rank order of the item information and by reducing a ratio scale measure to an ordinal scale measure, ARI loses the actual shape of the distribution (i.e., ARI changes a skewed distribution to a uniform distribution) and therefore ARI is, in theory, less optimal than ASI due to the information loss.

We considered two different weighting schemes to assign weights on the two information pieces when constructing an aggregate index, both theoretical and empirical weights generate more accurate results than simple balanced weights in most simulation conditions. While theoretical weights are easier to implement because it is simply a linear function of test length, empirical weights need to be adjusted through trial-and-error process until good upper bounds are selected for both information pieces. In addition, we also propose an attribute-level weight that is flexible enough to accommodate different weights for different attributes; when putting more weight on the attribute that is less accurately estimated, the recovery rate of that attribute and, in some cases, the whole pattern increases. The attribute-level weighting method provides a possibility if the accurate estimation of a particular set of attributes is considered more important in the testing practice. Because the attribute-level weight vector, $(w_1, w_2, ...w_M)$, is multiplied only to $KL(\hat{\boldsymbol{\alpha}})$ or $PWKL(\hat{\boldsymbol{\alpha}})$, this

proposed weighing scheme can be applied in traditional cognitive diagnostic CAT (CD-CAT, i.e., Cheng, 2009) as well.

In sum, the proposed approach is built upon two separate models, assuming both the unidimensional IRT model and the CDM fit adequately well to a single data set. As we have demonstrated, when the actual responses are generated from the HO-DINA model with high correlations among the attributes, or when the attributes display a linear structure, this assumption is satisfied. One might argue that using "misspecified" model is statistically suboptimal, but it is practically more appealing given the fact that practitioners would be able to calibrate two separate models much easier. It would be interesting as future research to compare the proposed dual-purpose approach with the adaptive test built directly upon the HO-DINA model. Because the HO-DINA model is a constrained version of the DINA model, the item selection methods for the HO-DINA model would be the same as those proposed for traditional CD-CAT, with the exception that the interim $\hat{\boldsymbol{\alpha}}$ estimate needs to be adjusted and $\hat{\theta}$ estimate needs to be included. As a result, our conjecture is the precision of $\hat{\boldsymbol{\alpha}}$ from HO-DINA-based CAT will be close to that obtained in our approach when $w = 1$, whereas the accuracy of $\hat{\theta}$ estimates will differ. Future simulation studies could be done to verify our hypothesis. In addition, further research should look at the performance of the method under multiple constraints such as item exposure and content balancing constraints. By incorporating these possible constraints, the proposed method will be able to satisfy most practical requirements in an educational testing setting.

The proposed method complements existing methods available for dual-purpose CAT, such as constraint weighted index (Wang et al., 2012) and shadow test method (McGlohen & Chang, 2008). Interested researchers can compare the performance of these three distinct approaches via simulation studies in terms of measurement precision and item exposure balance. Last but not least, we only consider fixed-length scenario in this study, but given the fact that examinees with more extreme ability levels might need longer test to have reliable ability estimates, variable-length CAT would be a promising future direction as well.

Note

[1]Researchers usually assume that the Q-matrix is constructed by subject matter experts and test developers (Cheng, 2009; McGlohen & Chang, 2008), although the development of the Q-matrix using these means has proven to be quite time-consuming and costly (Roussos, Templin, & Henson, 2007). Fortunately, de la Torre (2008) and J. Liu, Xu, and Ying (2013) have recently proposed to estimate the Q-matrix using statistical methods. As Q-matrix estimation is not the focus on the current paper, we assume it is known in advance.

Appendix A: The Conditions Under Which the Two Models Are Fitted Properly

To justify that both the DINA model and the IRT model can fit properly to the same response data, we conducted a small scale simulation study by varying the number of attributes (K = 3 and 8) and the correlation levels between attributes

(0, .5, .8). Examinees' θ were generated from the standard normal distribution (sample size is 5,000), λ_k and λ_{0k} were chosen to produce desirable correlations. Slipping and guessing parameters were generated from Uniform (0, .2), $\boldsymbol{\alpha}$'s and responses were generated from the HO-DINA model. Test length was set to be 100. Given the observed responses, we retrofitted the data with both the DINA model and the 2PL model separately. The correlations between the estimated $\hat{\theta}$ from the 2PL model and the θ from the HO-DINA model were computed, so were the recovery rates of $\hat{\boldsymbol{\alpha}}$, as well as the root mean squared error (RMSE) of slipping and guessing parameters, as shown in Tables A1 and A2.

Clearly, the data generated from the HO-DINA model can be retrofitted nearly perfectly by the DINA model, with both attribute and pattern recovery rate over .9, and RMSE of slipping and guessing parameters close to 0. The consequence of model misspecification is negligible. This is because the HO-DINA model can be viewed as a special case of the DINA model. We also present the observed (i.e., generated from the HO-DINA model) versus model predicted (retrofitted from the DINA model) total score distribution in Figure A1 for the high correlation condition, and the discrepancy is almost invisible. Results from medium and no correlation conditions are similar and omitted here.

On the other hand, if the correlations among the attributes are high, then the 2PL model can also be fitted properly with the resulting $\rho_{\theta\hat{\theta}}$ over .8, regardless of the number of attributes. This conclusion is consistent with de la Torre and Douglas (2004), in which they find that θ estimated from the HO-DINA model correlates highly (the correlation coefficient equal to .96 for the fraction-subtraction data) with

Table A1
Retrofit Data (From the HO-DINA Model) by the DINA and 2PL Model When K = 3

Correlation Among Attributes	RMSE			Attribute Recovery			Pattern Recovery
	$\rho_{\theta\hat{\theta}}$	Slipping	Guessing	At1	At2	At3	
.0	.00	.01	.00	1	1	.999	.999
.5	.69	.00	.01	1	1	1	1
.8	.80	.00	.00	1	1	1	1

Table A2
Retrofit Data (From the HO-DINA Model) by the DINA and 2PL Model When K = 8

Correlation Among Attributes	MSE			Attribute Recovery Rate								Pattern Recovery Rate
	$\rho_{\theta\hat{\theta}}$	Slipping	Guessing	At1	At2	At3	At4	At5	At6	At7	At8	
.0	.00	.00	.00	.99	.98	.99	.97	.98	.98	.98	.98	.92
.5	.79	.01	.00	.99	.99	.99	.99	.99	.99	.99	.99	.95
.8	.81	.00	.01	.99	.98	.97	.98	.93	.99	.99	.99	.86

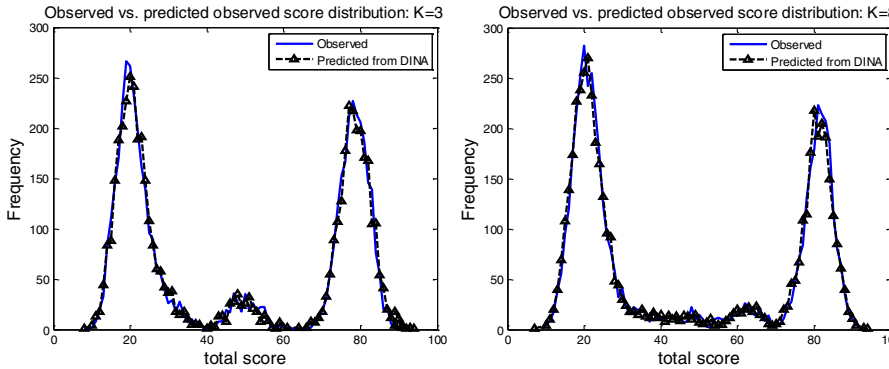

Figure A1. Observed (generated from the HO-DINA model) versus model predicted (retrofitted with the DINA model) total score distribution (high correlation).

ability estimate $\hat{\theta}$ obtained from a 2PL item response model. In addition, the correlation between the 2PL model predicted and observed proportion correct is as high as .99. Based on the results, it is legitimate to fit the response data with two separate models if the responses come from the HO-DINA model with high correlations among attributes.

Appendix B: Model Fit Checking Procedure and Results for Simulation Study 1

In this appendix, we implemented our practical guideline and showed, step by step, how to check the fit of both the 2PL and the DINA models for a given simulated response matrix. As to the 2PL model fit, the observed and model predicted proportion correct is presented in Figure B1*a* for all 500 items, and the correlation between them is as high as .998, indicating that the item level proportion-correct is recovered well. In addition, the log-odds ratio was computed for any pair of items, *j* and *j*', and the absolute difference between the observed and model predicted odds ratio serves as an indicator of model fit based on the second moment (Chen et al., 2013), i.e.,

$$l_{jj'} = \left| \log\left(\frac{N_{11}N_{00}}{N_{01}N_{10}}\right) - \log\left(\frac{\hat{N}_{11}\hat{N}_{00}}{\hat{N}_{01}\hat{N}_{10}}\right) \right|. \tag{B1}$$

In Equation B1, N_{11} denotes the observed number of individuals who answered both items correctly, and $\hat{N}_{11}$ denotes the corresponding model prediction. The approximate standard error of $l_{jj'}$, $SE(l_{jj'})$, was computed as $SE(l_{jj'}) = \sqrt{\hat{N}(\frac{1}{\hat{N}_{11}} + \frac{1}{\hat{N}_{00}} + \frac{1}{\hat{N}_{10}} + \frac{1}{\hat{N}_{01}})/N}$. As a result, the Z-score was computed to test whether $l_{jj'}$ is significantly different from 0. Figure B1*b* shows the Z-score based on the log-odds ratio statistic, and it appears that most of the Z-scores are very close to 0, and none of them reached statistical significance (this is because the Z-score based on log odds-ratio was shown to yield extremely low Type I error rate; Chen

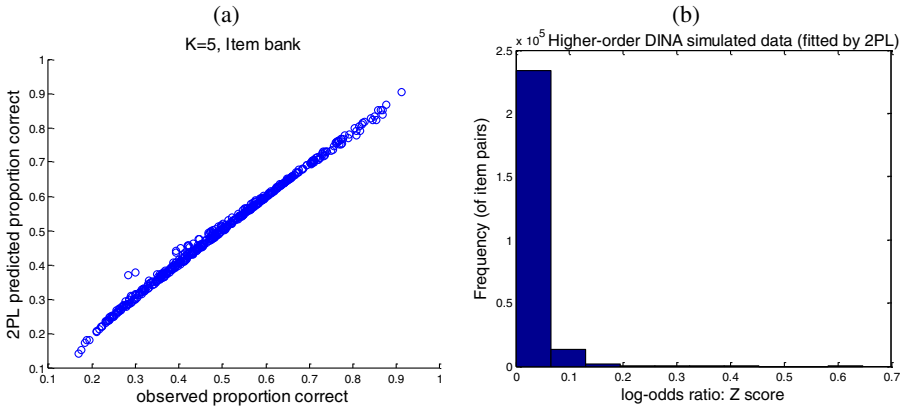

Figure B1. Model fit checking for the 2PL model.

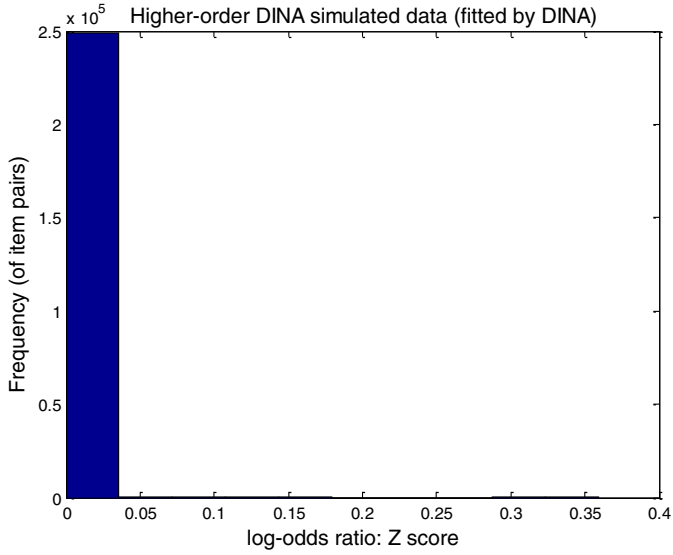

Figure B2. Model fit checking for the DINA model.

et al., 2013). Last but not least, we checked the item level fit using Yen's (1981) Q*1* statistic, which takes the form of

$$Q_{1j} = \sum_{l=1}^{L_q} N_l \frac{(O_{jl} - E_{jl})^2}{E_{jl}(N_l - E_{jl})}. \tag{B2}$$

Assume the examinees are classified into one of the L_q ability groups based on their estimated $\hat{\theta}$, this statistic quantifies the discrepancy between E_{lj} (expected number correct for item j and ability group l) and O_{lj} (the observed number of examinees belong to ability group l who answer item j correctly). With the 2PL model, Q_{1j} was shown to follow approximately a chi-square distribution with $L_q - 2$ degrees of freedom. Our analysis showed that out of 500 items, there were 42 (around 8.4%) items having a p-value (based on $Q1$) smaller than .05. Even though the detection rate was slightly over nominal level of 5%, the results reinforced that the 2PL model fitted the response matrix generated from the HO-DINA model adequately.

Fit analysis was also conducted for the DINA model fitting. Based on Chen et al.'s (2013) recommendation, we used the same Equation B1 and the Z-score of $l_{jj'}$ is plotted in Figure B2. It is shown that all Z-scores is kept below .8, indicating that none of $l_{jj'}$'s is significantly different from 0.

Appendix C: Descriptive Statistics of Item Bank and Examinee Sample in Studies 1 and 2

Table C1
Number of Items Measuring (or Examinees Mastering) Each Attribute

	Attributes				
	1	2	3	4	5
Number of items	234	250	257	235	237
Number of examinees (high correlation)	1,961	1,897	1,839	1,452	1,535

Table C2
Descriptive Statistics of Item Parameters

Item Parameter	a	b	Slipping	Guessing
Mean	.965	–.076	.115	.189
SD	.490	.705	.089	.148

Table C3
Item Parameters for Study 2

	a	b	c	s	g
Mean	.750	.049	.123	.256	.353
SD	.403	1.276	.048	.178	.112

Table C4
Q-Matrix Structure in the Test and Examinees Mastery Profile Distribution for Study 2

		At1	At2	At3	At4	At5	At6	At7	At 8
Number of items measuring each dimension		57	45	80	60	40	20	40	32
Number of examinees mastering each dimension		1,628	1,406	1,573	686	315	1,239	599	471
Total number of attributes	0	1	2	3	4	5	7	7	8
Number of items measuring certain number of dimensions	0	330	22	0	0	0	0	0	0
Number of examinees mastering certain number of dimensions	372	55	167	167	553	87	128	156	315

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*, 493–513.

Chang, H. H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Ed.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 195–226). Charlotte, NC: Information Age.

Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

Chang, H. H., & Ying, Z. L. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222.

Chen, J. S., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnostic modeling. *Journal of Educational Measurement*, *50*, 123–140.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632.

Cheng, Y., & Chang, H. H. (2007, April). *Dual information method in cognitive diagnostic computerized adaptive testing*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369–383.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 504–517.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 312–336). New York, NY: Wiley.

Hartz, S. M. C. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation), University of Illinois, Urbana-Champaign, IL.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407–419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3–16.

Liu, H., You, X., Wang, W., Ding, S., & 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.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 609–618.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808–821.

Partnerships for Assessment of Readiness for College and Careers. (2013). *PARCC assessment design*. Retrieved from http://www.parcconline.org/parcc-assessment-design

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*, 293–311.

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

Smarter Balanced Assessment Consortium. (2013). *The Smarter Balanced Assessment*. Retrieved from http://www.smarterbalanced.org/smarter-balanced-assessments/

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331–352.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the 27th annual meeting of the Military Testing Association, San Diego, CA.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *51*, 337–350.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55–73.

Tatsuoka, K. K. (1991). *Boolean algebra applied to determination of universal set of knowledge states (Tech. Rep. RR-91-44-ONR)*. Princeton, NJ: Educational Testing Service.

U.S. Department of Education. (2009). *Race to the top program executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, *20*, 373–388.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588.

von Davier, M. (2013, July). *Attributes, model equivalencies, hierarchies, and labels*. Paper presented at 2013 International Meeting of Psychometric Society, Arnhem, The Netherlands.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional "diagnostic" classification models: A commentary. *Psychometrika*, *79*, 340–346.

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.

Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*, 95–109.

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*, 255–273.

Xu, X., Chang, H., & Douglas, J. (2003, April). *Computerized adaptive testing strategies for cognitive diagnosis.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.

Authors

CHUN WANG is Assistant Professor, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN, 55401; wang4066@umn.edu. Her primary research interests include item response theory, computerized adaptive testing, response time modeling, and cognitive diagnostic modeling.

CHANJIN ZHENG is a PhD student of Educational Psychology at the University of Illinois at Urbana-Champaign, 210 Education Building, 1310 S. Sixth Street, Champaign, IL 61820; czheng5@illinois.edu. His primary research interests include item response theory models, cognitive diagnostic models, and their applications in computerized adaptive testing (CAT and CD-CAT).

HUA-HUA CHANG is Professor of Psychology, Educational Psychology, and Statistics, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL, 61820; hhchang@illinois.edu. His research interests are broad, encompassing theoretical development and applied methodologies, including computerized testing, statistically detecting biased items, cognitive diagnosis, and asymptotic properties in item response theory.